2021

Missing Pieces in Health Services Cost Analysis: Consensus on Modeling, Magnitude, and Micro-Costing

Mary Judith Dooley
Medical University of South Carolina

MISSING PIECES IN HEALTH SERVICES COST ANALYSIS: CONSENSUS ON MODELING, MAGNITUDE, AND MICRO-COSTING

BY

Mary Judith Dooley

A dissertation submitted to the faculty of the Medical University of South Carolina in partial fulfillment of the requirements for the degree
Doctor of Philosophy
in the College of Health Professions

MISSING PIECES IN HEALTH SERVICES COST ANALYSIS: CONSENSUS ON MODELING, MAGNITUDE, AND MICRO-COSTING

BY

Mary Judith Dooley

Approved by:

Chair, Project Committee	Kit N. Simpson, Dr.PH	Date
Member, Project Committee	Annie N. Simpson, Ph.D.	Date
Member, Project Committee	Paul J. Nietert, Ph.D.	Date
Member, Project Committee	J. Dunc Williams, Jr., Ph.D.	Date
Dean, College of Health Professions	Zoher F. Kapasi, Ph.D., PT, MBA	Date

Abstract of Dissertation Presented to the
Doctor of Philosophy Program in Health and Rehabilitation Science
Medical University of South Carolina
In Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy

MISSING PIECES IN COST ANALYSIS:
CONSENSUS ON MODELING, MAGNITUDE, AND MICRO-COSTING

By

Mary Judith Dooley

Chairperson: Kit N. Simpson, Dr.PH.
Committee: Annie N. Simpson, Ph.D.
Paul J. Nietert, Ph.D.
J. Dunc Williams, Jr., Ph.D.

Cost and cost savings have become an important focus for health policy administrators. However, there are missing pieces in our approach to cost analysis; there is no consensus on multivariable methods, no indicators of minimally acceptable values, and no specification of process costing.

In this dissertation, I propose to fill the gaps in the literature by 1) identifying which methods are appropriate for large claims data, 2) examine existing methods to establish minimally important difference (MID) in health outcomes to identify MID in costs, and 3) determine differences in sick visit clinic costs using a modified micro-costing method.

Most models that were compared to the generalized linear models Gamma distribution with log link found it to be the superior model in both simulated data and real administrative data. We recommend that in cases where acceptable anchors are not available to establish an MID, both the Delphi and the distribution-method of MID for costs be explored for convergence. Our micro-costing approach is feasible to use under virtual working conditions; requires minimal provider time; and generates detailed cost estimates that have “face validity” with providers and are relevant for economic evaluation.

Acknowledgements

I would like to thank my mentor and dissertation chair, Dr. Kit Simpson, for all her guidance and knowledge in the areas of cost analysis and health services research. Her ability to explain concepts when I did not follow and her willingness to meet when I got stuck was instrumental in both my gained knowledge and success. Her positive attitude was always encouraging, especially her ability to turn what initially may be felt as a failure or wasted time to, instead, gained knowledge and experience. Never wasted. Her reminder that a good dissertation is a done dissertation! Her mentorship continues to be a great source of guidance and knowledge professionally.

I would like to thank Dr. Paul Nietert for providing guidance as a mentor during my studies in biostatistics and committee chair for my master's degree to, ultimately, his expertise as a member of my dissertation committee assuring my work was statistically sound.

A special thank you to Dr. Annie Simpson, for introducing me to the doctoral program in College of Health Professions and serving as a member of my dissertation committee. For all the guidance provided in many areas of health services research and statistics over an undisclosed long period of time.

I would like to thank Dr. Dunc Williams for serving on my committee and all his feedback and input from a financial perspective into my dissertation. His comments helped to shape the conclusions to address the impact the results may have from various perspectives beyond the perspective of just my given method.

I would also like to thank Dr. Martina Mueller for being a great mentor and boss during my time working in the College of Nursing. Her support and understanding extended across both professional and academic endeavors.

Finally, I would like to thank all the love and support received from my family. Especially my mom, who was there to help me balance the work/school/family life when I was limited to only 24 hours in a day. I would like to thank my sons, Tyler and Ethan, for all the time with me they sacrified while I chased my goals and for being proud of me when I finished. Lastly, thank you to my sister, Rachel and brother, Fredrick, for allowing me to cry and vent when needed and providing me with encouragement.

Table of Contents

TABLE OF FIGURES

TABLE OF TABLES

CHAPTER 1

INTRODUCTION

The United States (US) has one of the highest annual growth rates for healthcare spending in the world, outpacing the gross domestic product (GDP) deflator by almost 5%, indicating that the increase in health care costs are more than what can be accounted for by inflation (Kaiser Family Foundation, 2011; Claxton, Rae, Levitt, & Cox, 2018; Zilberberg & Shorr, 2010). Concern about the growing spending trend in healthcare has prompted clinical and health policy decision makers to continually assess benefits and value of new treatments and care processes with an objective to control costs without sacrificing the quality of care. Tasked with the need to efficiently allocate limited resources, decision makers must consider the impact of money spent on one resource compared to the benefit that same money could achieve if it were spent on other resources (Zilberberg & Shorr, 2010; Weinstein, Siegel, Gold, Kamlet, & Russell, 1996; Gammon & Cotten, 2016; Tan, Rutten, van Ineveld, Redekop, & Roijen, 2009; Sanders, et al., 2016; Shander, et al., 2010).

The US is increasingly focusing on value of care, of which cost is an essential component. However, there are missing pieces in our approach to cost analysis; there is no consensus on multivariable methods, no indicators of minimally acceptable values, and no specification of process costing.

1.1. BACKGROUND

1.1.1 Multivariable methods in cost analysis

Cost and cost savings are becoming increasingly important for the US healthcare system. During the first 5 months of 2019, over one hundred papers with a key word of “cost savings” were published in major US journals. This overwhelming focus on cost savings makes it imperative that we begin to standardize our approach to cost analysis, because different approaches have different underlying statistical assumptions and often result in different outcomes.

Appropriate health care cost estimation is crucial as it is used to guide evidence-based health policy implementation. Health policy makers rely on costs to drive their decisions (Power & Eisenberg, 1998). The estimation for costs associated with a disease, such as diabetes, can influence the allocation of resources for the prevention and treatment of the disease (Fukuda, Ikeda, Shiroiwa, & Fukuda, 2016). As numerous administrative data sources have become available for analysis, it is essential that these data are analyzed properly. These claims data, or billing data, are usually observational data sources, such as Medicare billing data, that are often assessed for health care cost outcomes. Inaccurate cost analysis can lead policy makers to make sub-optimal decisions.

1.1.2 Minimally acceptable values

The determination of a minimally acceptable difference for clinical measures can be easily assessed through repeated use and clinician experience from observations of the outcomes to identify what is clinically important. Unfortunately, for more subjective outcomes, such as quality of life, the determination of a minimally acceptable difference requires an interpretation that can be understood by clinicians to judge effect of magnitude.

Minimal clinically important difference (MCID) was developed by Jaeschke, Singer, Guyatt (Jaeschke, Singer, & Guyatt, 1989) to create interpretability of change in score of

Quality of Life (QOL) Questionnaires in the “most influential paper in MID history” (King, 2011). Jaeschke et al. defined MCID, or as later referred to as minimally important difference (MID), as the “the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient’s management” (Jaeschke, Singer, & Guyatt, 1989). There are three methods to determine MCID: anchor-based, distribution-based, and consensus-based.

MIDs have been established to be useful for planning studies using quality of life and clinical assessment tools. These areas use guiding rules for MID to judge effect magnitude, however, we have found no publications establishing a MID for costs.

1.1.3.1 Process costing of value of care

As healthcare costs increase at alarming rates, there is a need to have accurate information when making decisions based on the value (cost-effectiveness) of health interventions and health care processes. Decision makers must determine the most efficient allocation of limited resources while delivering the best quality of care. Typically cost analysis is evaluated within an organization using the hospital internal accounting system, a system that is developed to maximize profitability and may not be representative of the true costs of care (Carroll & Lord, 2016; Zilberberg & Shorr, 2010). Nonetheless, these costing results often guide the decision-making about health care process. Better costing methods are therefore needed to provide a more accurate cost analysis to make better informed decisions. Time-driven activity-based costing (TDABC) is a less frequently used, but more accurate, micro-costing methodology that identifies true costs using service specific activity and resource use evaluated with patient specific treatment times (Gammon & Cotten, 2016; Carroll & Lord, 2016; Tan, Rutten, van Ineveld, Redekop, & Roijen, 2009). The TDABC method has been used to identify areas for process improvement, though, it has not always been cost efficient to perform (Shander, et al.,

2010). The biggest barrier to implementing TDABC is that it is resource intensive; requiring research staff with expertise to know what to assess to be present in the clinic to observe and record the care processes of each patient using a stopwatch to manually collect the timing of processes and resource use.

1.1.3.2 Electronic Health Records and Value Measurement

Heath care systems have taken an interest in the secondary use of electronic health records (EHR) for data as a cost-effective alternative to evaluate quality and efficiency of processes. The use of EHR timestamps have been established as a feasible and valid source for accurate data (Wu, et al., 2017) and its versatility as a resource is demonstrated by the various ways recent studies have utilized its secondary use. EHR time logs have been used to assess efficiency in workflow models and identify areas for process improvement, as well as, evaluate quality of care processes and determine when care deviates from usual process of care (Zhang, Mehotra, Liebovitz, Gunter, & Malin, 2013; Karp, Freeman, Simpson, & Simpson, 2018; Chen, et al., 2015).

1.2. PROBLEM STATEMENT

The literature has demonstrated limited research in the approach to cost analysis. These gaps are evident in existing publications regarding costs. This project will establish better methods to address these gaps in the literature.

1.2.1 Problem Statement 1

The analysis of large observational datasets come with its own challenges. The data are often skewed and group comparisons often bias. Cost data are commonly positively skewed with a few outlying patients, often with complications or very ill, disproportionately consisting of much of the total costs (Dodd, Bassi, Bodger, & Williamson, 2006; Bohl, Blough, Fishman, Harris, & Phelan, 2013; Malehi, Pourmotahari, & Angali, 2015; Kurz, 2017). Administrative data can often include zero costs or zero observations (non-users) that may make up a large proportion of the data (Malehi,

Pourmotahari, & Angali, 2015; Kurz, 2017). Further challenges arise when analysis of cost data also need to consider censored data within the research time frame (Dodd, Bassi, Bodger, & Williamson, 2006). Healthcare cost can vary according to region, health care system, population, and payer; thus, adjustment of costs is needed for comparisons across these factors. There are many different methods currently being used to estimate costs including: generalized linear models with a log link, natural logarithm transformed costs, gamma distribution, two-part models, and Bayesian models. As cost drives many health care policies, inaccurate analyses of cost can have serious consequences. A review of the current state of the evidence is needed to determine which approach is appropriate and valid, without which we cannot judge the quality of approaches being used in current studies.

1.2.2 Problem Statement 2

Studies to assess costs are abounding but their design and planning is hampered by the lack of indicators of magnitude of an important effect size for costs incurred in different settings. Other areas, such as quality of life measurement, use guiding rules for minimally important difference (MID) to judge effect magnitude. MIDs are useful for planning studies using quality of life and clinical assessment tools. However, we have found no publications establishing a MID for costs.

1.2.3 Problem Statement 3

Better costing methods are needed to provide a more accurate cost analysis to make better informed decisions. Time-driven activity-based costing (TDABC) is a less frequently used, but more accurate, micro-costing methodology that more accurately identifies costs using service specific activity and resource use evaluated with patient specific treatment times (Gammon & Cotten, 2016; Carroll & Lord, 2016; Tan, Rutten, van Ineveld, Redekop, & Roijen, 2009). The TDABC method has been used to identify areas for process improvement, though, it has not always been cost efficient to perform

(Shander, et al., 2010). The biggest barrier to implementing TDABC is that it is resource intensive; requiring research staff present in the clinic to observe and record the care processes of each patient using a stopwatch to manually collect the timing of processes and resource use. A solution to this shortcoming could be through use of electronic health records (EHR), providing a more efficient approach to record times of the sequence of events via electronic time stamps to implement TDABC. EHR activity logs have been effectively utilized to evaluate quality and efficiency of workflow and process of care (Mans, et al., 2008; Chen, et al., 2015; Wu, et al., 2017).

1.3. RESEARCH HYPOTHESES

1.3.1 Research Question 1

What methods of cost analysis are statistically and mathematically appropriate to use with large claims data and is there one method that could be considered optimal?

1.3.2. Research Question 2

Can existing methods to determine minimally important differences in health outcomes be used to identify minimally important differences in costs?

1.3.3. Research Question 3

Will the pediatric sick visit weighted labor cost, estimated using a modified TDABC, for a clinic visit before the COVID-19 pandemic show a difference in mean cost than a clinic visit during the COVID-19 pandemic?

Aim 1

Use structured provider interviews to identify the major variations in patient flow and describe the care process and resource use using TDABC workflow diagrams with time indicators for a moderate complexity visit for an established patient.

Aim 2

Compare providers' minute estimates to minutes estimated from EHR time stamps from telehealth "dashboard" data, and from Clarity EHR data, and estimate the uncertainty in

minute values using distributional parameters derived from CMS CPT standards and Monte-Carlo simulations for 100,000 visits to estimate uncertainty in labor cost estimates.

Aim 3

Aggregate flow chart estimates and uncertainty measures in a decision tree and calculate visit labor costs for an in-person sick visit before COVID-19 to a sick visit during the COVID-19 pandemic.

H1: Mean weighted labor cost before COVID - Mean weighted labor cost during COVID is < the MID defined by the anchor method as a difference between mean payment for adjacent CPT codes

1.4. DEFINITION OF TERMS

US = United States

GDP = gross domestic product

TDABC = time-driven activity-based costing

PA = Physician assistant

EHR = Electronic health record

MID = minimally important difference

MCID = Minimal clinically important difference

QOL = Quality of Life

ES = Effect size

PRISMA = Preferred Reposting Items for Systematic Reviews and Meta-Analysis

OLS = ordinary least squares

GLM = generalized linear model

NLR = normal linear regression

RMSE = root mean square error

MAE = mean absolute error

EEE = extended estimating equations

2SLS = 2-stage least squares

FIMSL = full information maximum simulated likelihood

LATE = local average treatment effect

ATET = average treatment effect on the treated

GAMLSS = generalized additive models for location scale and shape

MAPE = mean absolute predication error

MSE = mean square error

FE = fixed effects

CVD = cardiovascular disease

SD = standard deviations

MUSC = Medical University of South Carolina

MD = medical doctor or physician

NP = nurse practitioner

TH = Telehealth

PRO = patient-reported outcomes

DRG = Diagnosis-Related Group

CHAPTER 2

LITERATURE REVIEW

The literature review will not be an extensive review as this project is demonstrating the limited research and gaps in the literature regarding the approach to cost analysis.

2.1. Multivariable methods in cost analysis

Cost and cost savings are becoming increasingly important for the US healthcare system. During the first 5 months of 2019, over one hundred papers with a key word of “cost savings” were published in major US journals; 7 in JAMA, 19 in PLoS ONE, and the balance in other journals aimed at decision makers. This overwhelming focus on cost savings makes it imperative that we begin to standardize our approach to cost analysis, because different approaches have different underlying statistical assumptions and often result in different outcomes.

Appropriate health care cost estimation is crucial as it is used to guide evidence-based health policy implementation. Health policy makers rely on costs to drive their decisions (Power & Eisenberg, 1998). The estimation for costs associated with a disease, such as diabetes, can influence the allocation of resources for the prevention and treatment of the disease (Fukuda, Ikeda, Shiroiwa, & Fukuda, 2016). As numerous administrative data sources have become available for analysis, it is essential that these data are analyzed properly. These claims data, or billing data, are usually observational

data sources, such as Medicare billing data, that are often assessed for health care cost outcomes. Inaccurate cost analysis can lead policy makers to make sub-optimal decisions.

Analysis of large datasets of observational data come with its own challenges. Data are often skewed and group comparisons often biased. Cost data are commonly positively skewed with a few outlying patients, often those with complications or who are very ill, disproportionately consisting of much of the total costs (Dodd, Bassi, Bodger, & Williamson, 2006; Bohl, Blough, Fishman, Harris, & Phelan, 2013; Malehi, Pourmotahari, & Angali, 2015; Kurz, 2017). Administrative data can often include zero costs or zero observations (non-users) that may make up a large proportion of the data (Malehi, Pourmotahari, & Angali, 2015; Kurz, 2017). Further challenges arise when analysis of cost data also need to consider censored data within the research time frame (Dodd, Bassi, Bodger, & Williamson, 2006). Healthcare cost can vary according to region, health care system, population, and payer; thus, adjustment of costs is needed for comparisons across these factors. There are many different methods currently being used to estimate costs including: generalized linear models with a log link, natural logarithm transformed costs, gamma distribution, median regression, two-part models, and Bayesian models. As cost drives many health care policies, inaccurate analyses of cost can have serious consequences.

A review of approaches that compare cost analysis methods, which are conducted statistically and mathematically, is essential to provide needed evidence as to which methods are the most appropriate and valid for the evaluation of claims data. Therefore, we conducted such a systematic review to identify what methods of cost analysis are statistically and mathematically appropriate to use with large claims data and, specifically, determine whether one method could be considered optimal. The knowledge gained from

this review can be used to guide evidence-based cost analysis and properly assist policy makers' decisions.

The systematic review aimed to identify what methods of cost analysis are statistically and mathematically appropriate to use with large claims data. While most papers used established methods, there were three papers that introduced new methods. The most commonly assessed models were OLS and Gamma distribution models. The GLM Gamma distribution with Log link performed most consistently as the superior model in comparisons using both simulated data and real administrative data. The literature review suggests that this is the most appropriate model to use with administrative data. Some caution is suggested when dealing with heteroscedastic data or data with high proportion of zero costs (or non-users). The Tweedie distribution is an emerging new method that may be useful in future research (Kurz C. , 2017).

2.2. Minimally Important Difference

Minimal clinically important difference (MCID) was developed by Jaeschke, Singer, Guyatt (Jaeschke, Singer, & Guyatt, 1989) to create interpretability of change in score of Quality of Life (QOL) Questionnaires in "most influential paper in MID history" (King, 2011). Jaeschke et al. defined MCID, or as later referred to as minimally important difference (MID), as the "the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management" (Jaeschke, Singer, & Guyatt, 1989). There are three methods to determine MCID: anchor-based, distribution-based, and consensus-based.

The anchor-based method assesses the relationship between the change in score of the inconclusive questionnaire (target) with an independent measure (anchor) with an already known interpretation and association with the target (Gyatt, et al., 2002). These anchors are either individual based, a single anchor that uses within-patient or between-

patients anchors for interpretation, or population based, a multiple anchor that uses population based clinical measures for interpretations. For the latter, the change scores in the target QOL are compared to the proportion of people experiencing a clinical measure (e.g. health utilization, suicidal ideation, walking a block, job loss, mortality) (Gyatt, et al., 2002). Unfortunately, the multiple anchor approach does not have a lot of information yet, and no examples of specific studies were discussed.

The single anchor seeks to quantify the changes in QOL into trivial, small, moderate, or large categories. There are two methods used to determine these categories: with-in patient and between-patient. The within-patient method uses a patient's own reported change from their perspective as reference. The target QOL questionnaire is given at each timepoint, including baseline, and a global rating questionnaire additionally completed at all subsequent visits (not baseline). The global rating questionnaire asks the patient to rate if their "condition" has improved (e.g. no change, better, or worse) from their perspective since the last visit. Any change is then quantified by the patient on a 7-point Likert scale (e.g. 1. Almost the same, 2. A little worse/better, 3. Somewhat worse/better, 4. Moderately worse/better, 5. A good deal worse/better, 6. A great deal worse/better, 7. A very great deal worse/better) and degrees commonly established as: 1-3 considered small changes (MID), 4-5 moderate changes, and 6-7 large changes. The MID range can then be determined by comparing the change in QOL target score for each of the degrees (i.e. small, moderate, and large) and using ROC curves to minimize the misclassifications for optimal cut-points (Gyatt, et al., 2002). The between-patient method is similar to the with-in patient method. A global ratings questionnaire is completed the target questionnaire at all subsequent visits. However, instead of comparing their status to their previous status, patients pair off to discuss their statuses and then rate their status compared with their partnered pairs' status (same, worse, better). The MID range is

determined by the difference in QOL score in patients that rate themselves “little better” or “little worse.”

There are several limitations to the single anchor method. The same absolute difference in score may have different meaning across different portions of the scale (e.g. 20 to 30 mean something different than 90 to 100) and Gyatt believes the proportion of patients achieving benefit is more important than mean difference (Gyatt, et al., 2002). There is also the potential for recall bias, for example, the prospective of change is correlated with present condition (King, 2011; McGlothlin & R, 2014) and the sample size in each degree of global rating change group may be small (King, 2011).

The distribution-based method of determining MID examines the relationship between the magnitude of effect and variability (Gyatt, et al., 2002). Typically, this is expressed as a ratio called Cohen’s D, where the magnitude is within patient difference and variability is between patient variability for the control group (baseline) or the pooled variability of control and treatment groups (baseline) (Gyatt, et al., 2002; King, 2011). There are 2 inherent limitations to this method to be considered. First, that variability is different for each study, thus, effect sizes may not be comparable across different populations with varying degrees of homogeneity (Gyatt, et al., 2002). The second limitation is the interpretability of effect size in terms of standard deviation is not easily understood and applicable to practicing clinicians. Cohen sought to address the latter limitation by suggesting ranges of 0.2. 0.5, and 0.8 as small, moderate, and large changes, respectively. There has been some concern as to the arbitrariness of these cut points; however, studies have provided evidence that suggest the plausibility of Cohen’s ranges and consistency of standard deviations and MID within a same instrument.

The final method, consensus (Delphi), was developed by the Rand Corporation around 1950. Expert judgement is often needed to solve complex problems when a definitive conclusion is not available. The Delphi uses expert opinion in a series of rounds

(King, 2011). Its first significant use was in 1953, by Dalkey and Helmer, to assess 7 experts' judgement of atomic warfare (Pill, 1971; de Villiers, de Villiers, & Kent, 2005; Okoli & Pawlowski, 2004). The objective of the method is to gain the judgement of a panel of selected experts in a field (Okoli & Pawlowski, 2004). The Delphi method is conducted anonymously using questionnaires sent by mail, e-mail, or fax; this anonymity reduces bias that may occur from dominant personalities within a group of experts (Okoli & Pawlowski, 2004). Each question provides an opportunity to provide feedback to explain choices while providing opportunities to for individuals to reevaluate their stance given information provided by other experts (Okoli & Pawlowski, 2004).

The Delphi method requires identifying and recruiting qualified experts; however, the method of eliciting responses by paper provides the opportunity to build a panel of experts from various geographical regions. The first questionnaire may be developed from a literature review conducted by the research team or created with open-ended questions with the intent to obtain expert opinion (de Villiers, de Villiers, & Kent, 2005). Any structured items are developed for the questionnaire using one of two methods to formulate the responses: ranking or rating. The ranking method requires the researchers to devise a ranking system for the experts to utilize, whereas, the rating system uses level of agreement from finite Likert scales (de Villiers, de Villiers, & Kent, 2005). Ranking system asks participant to rank a group of factors in order of priority. Likert scales are recommended to not include a neutral option to prevent participants from not providing a definitive rate, such as a 4-point Likert using strongly agree, agree, disagree, and strongly disagree (de Villiers, de Villiers, & Kent, 2005).

The first round of questionnaires sent out to the experts with structured items for ranking/rating as well as a comment explaining the decision for the individual expert's response (de Villiers, de Villiers, & Kent, 2005). Finally, experts are asked to provide overall feedback and questionnaires sent back to the researchers. The second round of

questionnaires is formulated by the analysis of the first round, with a summary sent back to each expert containing each item not meeting consensus including their original response as well as a summary of the other experts' responses. The review of opposing and similar opinions gives each participant a chance to reevaluate their response.

To analyze the items on the questionnaire requires various methods. The assessment of rankings is best analyzed using Kendall's W coefficient of concordance (Okoli & Pawlowski, 2004). Coefficients of 0.7 or higher correlates to a strong agreement and consensus is achieved, however, coefficients less than 0.7 require reevaluation by the experts (Okoli & Pawlowski, 2004). This is continued at each round until either the coefficient reaches 0.7, the maximum number of promised rounds are reached, or there are no significant change in rankings of successive rounds indicating a lack of movement toward consensus (Okoli & Pawlowski, 2004). For the assessment of ratings mean scores that fall in the middle, e.g. around 2.5 on a 4-point Likert scale, correspond to non-consensus (de Villiers, de Villiers, & Kent, 2005). Consensus is considered reached if rating score ceases to change significantly between rounds. The main limitation of this method is the lack of patient perspective.

2.3. Effect size in healthcare

Potential papers were identified by searching the database Scopus using the following search terms: "effect size" and "healthcare" in the title, abstract, or keywords. Searches were conducted for all articles published up to the end of January 2019.

Only papers reporting effect sizes were included, excluding those that define effect sizes as ratios (e.g. odds ratio, hazard ratio, incident rate ratios), correlations, or regression coefficients. Additional exclusions included papers not being published in English, method papers, book chapters, review papers, and those reporting effect size but giving no interpretation of the effect size. Data elements extracted from eligible papers include: paper's first author, year of publication, outcome(s) being measured by effect size,

effect size method(s) used, interpretation of effect size values or ranges, and citations used for effect size methods and/or interpretation. The titles and abstracts of all papers found utilizing the search terms in the database were reviewed for potential eligibility. Among those titles and abstracts meeting potential eligibility, the full text article was reviewed for eligibility to be included in the literature review.

The search of the database identified 698 potential papers, after title and abstract review there were 340 papers that met initial inclusion criteria for full article review (Figure 1). There were 184 papers that were included in the review and summary data extracted (Table 1). There were about 30 different, although many similar, definitions of effect size (ES) reported across the 184 papers. The most common type of ES calculation used was Cohen's d (49.7%), another 24 (13%) of papers provided the formula for ES used that was similar to Cohen's d, and about 10% did not define the type or provide a formula of ES they used. The interpretation of ES by ranges or cut points were not provide by some papers, offering only the interpretation of specific ES found, however, many papers did provide the ES ranges and cut points for interpretation. The most common cut points for the interpretation of small, medium/moderate, and large ES were 0.02 (43.2%), 0.5 (38.8%), and 0.8 (43.7%), respectively.

References regarding ES calculation and ES interpretation were examined across all the included papers with almost half (42.4%) not including a citation in the paper. Cohen was most frequently referenced author (41.9%; 72.6% of papers with references to ES) with over a third of the papers (34.8%) referencing Cohen's book "Statistical Power Analysis for Behavioral Sciences."

2.4. Develop an efficient tool for measuring costs for the calculation of value of care

Hospitals commonly use mean values based on data from their own accounting system when estimating the cost of care for decision making processes as this is the least resource intensive approach. The hospital accounting cost are often based on the internal

cost from accounting data with a “stepdown” allocation of indirect costs, where departments are ordered in a hierarchical fashion and all indirect costs are distributed step-wise to the departments below until only one department is left. The hospital accounting system depends on the institution’s historical production flows and may be skewed by factors associated with overhead allocations aiming at maximizing reimbursement and collection rates that best maximize profits. Thus, a hospital’s internal accounting system is skewed towards resource use patterns associated with the average patient and are often not representative of true cost (Zilberberg & Shorr, 2010; Carroll & Lord, 2016).

The true cost of a care process usually requires a micro-costing approach to measure. Micro costing establishes patient specific resource use directly attributed to actual diagnostics, devices, and drugs used in treatment and patient specific treatment times to determine actual labor resources use (Tan, Rutten, van Ineveld, Redekop, & Roijen, 2009; Carroll & Lord, 2016). This more detailed level of costing provides the ability to identify subpopulations (e.g. race, comorbidity, age group) most affected by costs (Tan, Rutten, van Ineveld, Redekop, & Roijen, 2009). Time-driven activity-based costing (TDABC) is a micro-costing methodology that identifies each specific activity and resource use with indirect costs estimated using hospital defined estimates for the specific resources used to determine true costs (Gammon & Cotten, 2016; Carroll & Lord, 2016). TDABC uses a management engineering approach of direct observation of the process and the collection of a large sample of relevant data by stopwatch and observation of the care process. Typically, TDABC is conducted with research staff present in the clinical setting during all operational hours to be immediately available when patients arrive for treatment, tracing and documenting the specific activities related to the treatment manually by use of a stopwatch. Each research staff can only trace one patient at a time, therefore, in the event that more than one patient is going through the process during overlapping

times, opportunity to collect data on the additional patients is lost. Furthermore, data needs to be collected on approximately 40-50 patients to have obtained a sample large enough to satisfy the statistical “law of large numbers” (more than 30 viable observations) needed to evaluate costs.

Literature suggests that TDABC is an effective method to find true costs which can be used to identify inefficiencies and cost drivers within processes. In one study, researchers demonstrated how this precise method of costing was used to address a concern that health care administrators were emphasizing blood transfusions over other alternatives based on a misconception of actual cost, affecting quality of care. The researchers determined the true costs of blood transfusions, using activity-based costing to include the costs of acquiring and administrating, were much more costly than the mean acquisition costs used by administrators to inform clinical process decisions (Shander, et al., 2010). At Boston Children’s Hospital, TDABC and process mapping identified an area to improve efficiency by the addition of a physician assistant (PA) that decreased patient waiting times and increased revenue (Kaplan & Witkowski, 2014). Similar results were found when a psychiatrist in Norway used the process to analyze the efficiency of his clinic and discovered a need to modify the care process; resulting in better outcomes and improved capacity (Kaplan & Witkowski, 2014). Another study in a preoperative assessment center applied TDABC methods to evaluate the two-phase implementation of a process improvement initiative that ultimately reduced process time by 33% and cut cost of care by almost half, without negatively impacting outcomes (French, et al., 2013). These examples are indicative of the potential for process improvement and cost savings. Nonetheless, since this method for determining cost is very resource intensive, it is rarely employed for hospital cost estimation. The biggest barrier to implementing TDABC is that it is resource intensive; requiring research staff present in the clinic to observe and record the care processes of each patient using a stop watch to manually collect the timing of

processes and resource use. Consequently, it is not known how much the method used for costing affect the calculation of value.

A solution to this shortcoming could be through use of structured interviews and electronic health records (EHR), providing a more efficient approach to record times of the sequence of events via electronic time stamps to implement TDABC. EHR activity logs have been effectively utilized to evaluate quality and efficiency of workflow and process of care (Mans, et al., 2008; Chen, et al., 2015; Wu, et al., 2017).

2.5. Electronic Health Records and Value Measurement

The Panel of Cost-Effectiveness in Health and Medicine, a group organized to develop standardization in cost-effective analyses, recommends micro costing methods for studies conducted within organizations (Weinstein, Siegel, Gold, Kamlet, & Russell, 1996). Unfortunately, micro costing is both costly and time-consuming and, despite its greater accuracy, this is the main barrier to its implementation (Tan, Rutten, van Ineveld, Redekop, & Roijen, 2009; Carroll & Lord, 2016). Every time a hospital care process is changed in a major way, decision makers must decide if the additional costs of micro costing, with its high resource use and slow time line, is justified by its more accurate cost data (Tan, Rutten, van Ineveld, Redekop, & Roijen, 2009). Thus, a faster, less resource-intensive costing method would greatly benefit hospital managements' ability to identify good value for money invested in process improvements. A potential solution may be found in the electronic health record (EHR)

Health care systems have increasingly focused on the secondary use of EHR for data to evaluate quality and efficiency of workflow processes as a cost-effective alternative. This secondary application was demonstrated in a pilot study that successfully used EHR time logs to model workflows that evaluate efficiency and determine areas for process improvement (Chen, et al., 2015). Wu et al. evaluated the feasibility and validity of EHR timestamps use in the assessment of workflows and determined the timestamps

were a feasible and valid source of accurate data for identifying health care processes and reduced behavioral bias associated with being observed (Wu, et al., 2017). EHR time logs have been used to assess efficiency in workflow models and identify areas for process improvement, as well as, evaluate quality of care processes and determine when care deviates from usual process of care (Zhang, Mehotra, Liebovitz, Gunter, & Malin, 2013; Karp, Freeman, Simpson, & Simpson, 2018; Chen, et al., 2015). The use of computerized time stamps from EHR to validate the times of a sequence of events in the process of care has the potential for being a cost saving and accurate costing method. This approach is promising in that it significantly reduces the resource cost associated with research staffing needed to observe and record time segments in the process using a stopwatch.

2.6. Examples of studies that use less than optimal approaches to report on cost data

Cost data comes with its own challenges that often include non-negative values, large of number of zero costs (non-users of healthcare utilization), right skewed from a few outlying patients that disproportionately consist of much of the total costs, and heteroscedasticity (non-constant variance). Popular models in the literature to analyze these cost data are ordinary least squares (OLS) linear regression, with and without log transformation, and generalize linear models (GLM) with gamma distribution and log link, however, many studies comparing the two methods have found GLM gamma with log link to have superior performance over OLS regressions.

In 2003, Mandell et al. analyzed the service use and costs of psychiatric disorders using ordinary least squares (OLS) linear regression with no log transformation based on the robustness of large sample size (Mandell, Guevara, Rostain, & Hadley, 2003). While some state that as sample sizes get larger, the robustness of OLS with a log transformation improves similarly to GLM Gamma with Log link, however, it is not considered to be robust for heteroscedasticity and the authors did not use log

transformation (Malehi, Pourmotahari, & Angali, 2015). OLS without log transformation requires meeting assumptions of normality including homoscedasticity. Ignoring the violations leads to biased estimates that may provide incorrect conclusions (Dodd, Bassi, Bodger, & Williamson, 2006; Malehi, Pourmotahari, & Angali, 2015; Basu, Arondekar, & Rathouz, 2006). The GLM Gamma with Log link has been found a consistently much better model that provides more precise estimates than OLS and performs well with small bias for both small and large sample sizes. The GLM gamma retains the original scale of the data providing an accurate interpretation with the least amount of bias (Bohl, Blough, Fishman, Harris, & Phelan, 2013; Dodd, Bassi, Bodger, & Williamson, 2006; Malehi, Pourmotahari, & Angali, 2015; Manning & Mullahy, 2001).

Nichols et. al. analyzed the annual direct costs of follow-up medical utilization among patients that entered a cardiovascular disease (CVD) registry (Nichols, Bell, Pedula, & O'Keeffe-Rosetti, 2010). The authors stated they analyzed costs using Proc GLM with no transformation (i.e. OLS) for “straightforward interpretation of the parameter estimates.” The justification given for not transforming was they found “no change in direction or statistical significance of the results” when cost were analyzed using a log transformation. Finding a similar direction or statistical significance of costs in a known biased model by comparing it against a more widely accepted method does not validate the erroneous model’s cost estimates. OLS without transformation is shown to lead to bias estimates of mean costs that may provide incorrect conclusions; these conclusions can affect decisions based off those biased estimates (Dodd, Bassi, Bodger, & Williamson, 2006). The GLM Gamma with Log link is the best-fit model for highly skewed data, regardless of sample size, and provides accurate interpretation (Dodd, Bassi, Bodger, & Williamson, 2006; Bohl, Blough, Fishman, Harris, & Phelan, 2013; Malehi, Pourmotahari, & Angali, 2015).

Wang et al. examined hospitalization costs for stroke patients using regression analysis without log transformations (Wang, et al., 2014). The authors did not log

transform data based on the exclusion of 1% trim of the top and bottom of data to avoid outliers, large sample size, and the reference of 2 other published studies (discussed above) siting no transformation for ease of interpretation. The authors attempted to control the potential of outliers dominating model estimates and biasing results from heteroscedasticity of variance by excluding them instead of using a model that could handle these challenges. Large sample sizes may improve estimates in OLS models, though a threshold for size has not been determined to identify when it would no longer provide biased estimates (Dudley, et al., 1993). Studies comparing statistical modeling methods show that OLS with no transformation is prone to overestimate estimates which may impact decisions made from inaccurate estimates and conclusions.

In a more recent publication, Willink et al. published a cost-benefit analysis of hearing care services in 2019, using OLS (no transformation mentioned) to analyze total spending and spending by service type (Willink, Reed, & Lin, 2019). No rationale was given by the authors for not transforming the data or addressing violation of assumptions.

The GLM Gamma distribution with Log link has been found to be the most appropriate model for skewed data and/or heteroscedasticity, overall performed superiorly to OLS in simulated studies and real observational data with accurate estimates, and retains the data's original scale for accurate interpretations. Inaccurate and biased estimates can lead to incorrect conclusions that may impact decisions that affect the resource allocation in healthcare (Dudley, et al., 1993; Dodd, Bassi, Bodger, & Williamson, 2006; Basu, Arondekar, & Rathouz, 2006; Malehi, Pourmotahari, & Angali, 2015). OLS may perform well in data without heavy tails or if sample sized are large enough, the unbiased estimates are not consistent across studies. Furthermore, normal linear regression with heteroscedasticity may lead to negative prediction of cost (Dodd, Bassi, Bodger, & Williamson, 2006). While GLM Gamma distribution with Log link in not universally correct for all data, it has been shown in multiple model comparison studies to perform as well as

specialized models that fit specified data, allow for heteroscedasticity that is often found in cost data, be a good fit for highly skewed data, overall be most reliable regardless of sample size, and allow for the accurate interpretation of data on the original scale.

CHAPTER 3

METHODS

3.1. RESEARCH QUESTION 1

What methods of cost analysis are statistically and mathematically appropriate to use with large claims data and is there one method that could be considered optimal?

3.1.1 Study Design

A systematic review identified what methods of cost analysis are statistically and mathematically appropriate to use with large claims data. Only statistical method papers using multivariable modelling of cost, with or without methods controlling for selection bias, were included given they met one of the following inclusion criteria: 1) a comparison of two or more statistical methods to analyze cost or 2) one statistical method performed on two or more different types of cost data.

3.1.2 Data Collection

Data elements extracted from eligible papers included: paper's first author, year of publication, statistical methods being evaluated, data types used in analysis, year(s) the data was collected, simulation approach used, sample size of data, distribution of the data, the results of the method performance and the author's recommendations based on their results.

3.2. RESEARCH QUESTION 2

Can existing methods to determine minimally important differences in health outcomes be used to identify minimally important differences in costs?

3.2.1 Study Design

The project reported the MIDs derived using each of the three methods for cost data from a hospital admission cohort (high-costs) and a clinic visit cohort (low-cost).

3.2.2 Data Collection

The hospital admission cohort consists of patients identified as having an opioid-related event treated in any hospital in a state over a 3-year period. The clinic visit cohort consists of cost data from outpatient visits incurred over 12 months for HIV-infected adolescents from 4 clinics in different states in the US. These data were de-identified and are part of ongoing exploratory studies deemed non-human research by our IRB. The data are governed by data use agreements and not available for other use.

3.2.3 Data Analysis

MIDs were calculated using three methods: 1) anchor-based, 2) distribution-based, and 3) consensus-based. The anchor-based MID was calculated based on the relationship of clinic care costs with the 2017 Medicare medical fees for the median (50th percentile) cost for complex clinical visits. The distribution method was calculated based on Cohen's cutoffs of 0.2, 0.5, and 0.8 standard deviations (SD) for small, medium, and large effect size (ES), respectively. The consensus-based method was conducted by professionals from various backgrounds in an academic institution that assess cost evaluations through a questionnaire that was administered via email.

3.2.4 Limitations

MIDs are useful for planning studies using quality of life and clinical assessment tools. However, we have found no publications establishing a MID for costs.

3.3. RESEARCH QUESTION 3

Will the pediatric sick visit weighted labor cost, estimated using a modified TDABC, for a clinic visit before the COVID-19 pandemic show a difference in mean cost than a clinic visit during the COVID-19 pandemic?

Aim 1

Use structured provider interviews to identify the major variations in patient flow and describe the care process and resource use using TDABC workflow diagrams with time indicators for a moderate complexity visit for an established patient.

Aim 2

Compare providers' minute estimates to minutes estimated from EHR time stamps from telehealth "dashboard" data, and from Clarity EHR data, and estimate the uncertainty in minute values using distributional parameters derived from CMS CPT standards and Monte-Carlo simulations for 100,000 visits to estimate uncertainty in labor cost estimates.

Aim 3

Aggregate flow chart estimates and uncertainty measures in a decision tree and calculate visit labor costs for an in-person sick visit before COVID-19 to a sick visit during the COVID-19 pandemic.

> H1: Mean weighted labor cost before COVID - Mean weighted labor cost during COVID is < the MID defined by the anchor method as a difference between mean payment for adjacent CPT codes

3.3.1 Study Design

A mixed methods approach was used for data collection and analysis to perform a modified TDABC of a sick visit. A sick visit was defined as a low complexity clinic visit (CPT 99213), classified as a 15-minute face-to-face visit. Visits were described for children between the ages of 5-9 years old. The TDABC steps included: 1) recorded structured interviews with providers, 2) iterative workflow mapping, 3) EHR timestamps for time validation, 4) standard cost weights for wages, 5) clinic CPT billing code mix for complexity weights and 6) simulations to assess effects of uncertainty on cost differences.

3.3.2 Sample

This study was conducted using a pediatric clinic associated with an academic hospital in the southeastern United States. The Medical University of South Carolina (MUSC) is one of only two federally recognized telehealth Centers of Excellence, identifying it as having an established and successful telehealth program with high volume. MUSC has been providing telehealth programs across the state of South Carolina since 2005, reaching 42 of the 46 counties. Secondary data was obtained from the EPIC electronic health record system used by MUSC. The study met institution IRB definition of quality improvement project and did not require approval.

3.3.3 Data Collection

Structured interviews were used to collect data to map the care process of a sick clinic visit for an established patient age 5-9. An interview guide with five questions with probes was used (Appendix A). Interviews were conducted by two interviewers familiar with the workflow and EHR system. Two providers were interviewed separately: a physician (MD) and a nurse practitioner (NP). Both interviewers were present for both sessions and the interviews were recorded.

Two independent sources of data from the clinic were used to extract visit time stamps for clinic visits with CPT 99213. One set of process validation data was extracted from the EPIC telehealth dashboard, used by practice managers and telehealth personnel to monitor the processes in the clinic. Data were extracted for all clinic pediatric patients seen during September 2020 with a low complexity clinic visit (CPT 99213). These data included the timestamps for check-in, treatment start time, provider treatment team composition (i.e. MD or NP), record access by each actor, and printing timestamp used by providers to present a visit summary and care plan for patients at the end of the visit. A second data set was extracted from the EPIC Clarity Warehouse, which included all CPT 99213 clinic visits in September 2019 and September 2020.

To determine labor costs of each actor, median US salaries for each of the actor were established using the US Bureau of Labor Statistics salary data for 2019. Monte Carlo simulations were developed to mimic the variation of labor minutes by the providers and staff, thus the variation of total labor cost, in the clinic setting. The variation in minutes and cost per minute for actors provided a distribution of costs across 100,000 visits for in-person visits before COVID and telehealth and in-person visits during COVID.

3.3.4 Data Analysis

3.3.4.1. Aim 1

The recorded interviews were processed using Rapid Qualitative Analysis to develop the workflow charts. Iterative review of recording was used to reach agreement on clinic flow between interviewers. The resulting flow charts were reviewed and edited as needed by providers who then gave final approval of their position-relevant relevant charts. The clinic mapping process for patients utilizing the clinic for a sick visit, from signing-in at the beginning of the appointment to the conclusion of the clinic visit, began with the review of the recorded interviews. Identification of each step in the process was completed along with the determination of actors (e.g. MD, NP, nurse, front desk personnel) and approximate time in minutes to complete each step. Three workflow charts were then developed for 1) in-person clinic visits before COVID-19, 2) telehealth clinic visits during COVID-19, and 3) in-person clinic visits during COVID-19. For each of the workflow charts, steps of the process (identified by a square) are organized by the order in which they are completed. Potential additional or alternative steps are identified with decision nodes (diamond) in the flow chart. For each step of the process, the average time to complete is noted (contained within small circle in bottom right corner of squares), and actors involved (color-coded) are identified and listed. The three workflow charts were created from each interview, then the interviews were reviewed again to make edits to the

workflow charts. The two interviewers then met to review the workflow charts and both interviewers agreed on the construction of the workflow chart.

The workflow chart was reviewed while listening to the recorded interviews and check to be sure there were no missing connections and then returned to the interviewees for verification of accuracy. Once any suggested edits were completed, the interviewers reviewed the recorded interviews to ensure the accuracy of the workflow charts. This process was repeated as analysis of minutes and cost were conducted. The workflow charts generated by the MD and NP were combined to create a single clinic process workflow chart for each visit type before and during COVID-19.

Using time-driven activity-based costing methods, the true labor cost of care was assessed for in-person clinic visits before COVID-19 and the telehealth and in-person clinic visits during COVID-19. To determine labor costs of each actor, median US salaries for each of the actor were established using the US Bureau of Labor Statistics salary data for 2019. Total loaded salary for each actor was evaluated as median salary plus fringe benefits (e.g. health benefits, vacation package), the latter assessed at 35% of the median salary. A total of 2,080 annual hours worked were assumed for a full-time employee; a work capacity rate was assessed for nursing (e.g. Licensed Practical Nurse (LPN) and Certified Medical Assistant (CMA)) and administration staff at 80% (1664 hours) and for provider (e.g. MD and NP) at 72.3% (1504 hours). The cost per minute for each actor was calculated as the total loaded salary divided by the number of capacity hours per year divided by 60 (minutes). For process steps that were either completed by two different actors or may potentially be completed by two different actors, as identified in the workflow chart, a 50/50 weight was given to each actor's salary to estimate the cost per minute for the time in the mixed process step.

The cost of each actor for the visit is determined by the total minutes utilized multiplied by the cost per minute for the actor. For in-person and TH sick visit, the labor

costs across all actors are summed to determine the total labor cost of the clinic visit. Analysis of labor cost was conducted using Microsoft Excel.

3.3.4.2. Aim 2

For Aim 2, providers' minute estimates from interviews were validated by the minutes estimated from EHR time stamps from telehealth "dashboard" data, and from Clarity EHR data. Mean (SD) in minute values were used to validate the minute estimates in the clinic flow charts and to identify visits with CPT billing codes for tests and time stamps for validating prescription-related process effects. These data were used to estimate the mix of visits to generate complexity weights for the cost estimates. A second data set was extracted from the EPIC Clarity Warehouse, which included all CPT 99213 clinic visits in September 2019 and September 2020. These data were used to validate virtual visit time stamps and to estimate in-person visit time stamps to be used for estimating visit costs.

Monte-Carlo simulations were developed to estimate the uncertainty in minute values using distributional parameters derived from CMS CPT standards conducted with 100,000 simulated visits to estimate uncertainty in labor cost estimates. The provider time was simulated using a Weibull distribution identified by Medicare specified range for the visit type. All other staff time was modeled on a Beta PERT distribution defined by a minimum and maximum value. Median salary for actors were also varied on a normal distribution. The variation in minutes and cost per minute for actors provided a distribution of costs across 100,000 visits for in-person visits before COVID and telehealth and in-person visits during COVID. Simulations were conducted using Crystal Ball software.

3.3.4.3. Aim 3

Aim 3 aggregated flow chart estimates and uncertainty measures in a decision tree and calculated weighted sick visit labor costs for an in-person sick visit before COVID-19 to a weighted sick visit during the COVID-19 pandemic. EHR data extracted from EPIC

Clarity warehouse was used to identify and categorize providers to determine a provider mix for the same week in September for 2019 and 2020. Decision trees were constructed using the identified provider mix and delivery method mix to calculate an average weighted visit cost using the mean and +/- 1 standard deviation estimated from the labor cost Monte Carlos simulations conducted in Aim 2.

Differences were evaluated between weighted mean of forecasted visit cost estimated by simulations for before COVID and during COVID. Minimally important difference (MID) measured by a well-defined anchor has been identified as a conservative effect size for low-cost studies (Dooley, Simpson, Nietert, Williams Jr., & Simpson, 2021). The anchor-based MID was based on the relationship of clinic care costs between the low complexity sick visit (CPT 99213), defined as a 15-minute face-to-face clinic visit, and the moderate complexity sick visit (CPT 99214), defined as a 25-minute face-to-face clinic visit. The median Medicare medical fee in 2017 was $125 and $184, for low and moderate complexity visits, respectively. The meaningful payment difference between the two visits is $59 for 10 minutes.

3.3.5. Limitations

There is no direct benefit to patients analyzed for this study. The modified TDABC method will eliminate the need for research staff to trace patients to determine a workflow map in order to have the adequate sample size needed to conduct analyses. The results will be relevant for informing essential discussions about: 1) which TH programs to keep; 2) how to improve TH efficiency; and 3) least costly mixes of TH and in-person visits.

CHAPTER 4

MANUSCRIPT 1

Systematic review of statistical methods for analyzing healthcare cost in administrative data

4.1.0 Abstract

BACKGROUND: Healthcare costs are increasing at alarming rates in the United States (US) putting a heavy burden on the healthcare reimbursement system. Cost and cost savings have become an important focus as health policy administrators are tasked with determining the most effective allocation of limited resources. The availability of large databases, such as administrative data, comes with many challenges for analyses, including: skewed data, inflated zero counts, and potential selection bias among comparison groups. Thus, it is imperative that they are evaluated correctly. There are many different methods currently being used to estimate costs including: generalized linear models with a log link, natural logarithm transformed costs, gamma distribution, median regression, two-part models, and Bayesian models. This systematic review will identify which methods are statistically and mathematically appropriate for large claims data.

METHODS: Scopus and Ovid were searched for potential statistical method papers using multivariable modelling of cost that were published up to the end of February 2018. Inclusion criteria required either a comparison of two or more statistical methods to analyze cost or one statistical method performed on two or more different types of cost data. This systematic review follows the guidelines according to Preferred Reposting Items for Systematic Reviews and Meta-Analysis (PRISMA).

RESULTS: The review identified 1,048 potential papers, of which, 80 met the inclusion criteria for a full article review. There was a total of 9 papers included in the systematic review; one paper included simulations and eight papers assessed real cost data. There were 28 models assessed across the nine papers with ordinary least squares (OLS) and generalized linear models (GLM) being the most common.

CONCLUSIONS: GLM using the gamma distribution was included in all but two of the comparisons. Most other models that were compared to the GLM Gamma distribution with log link found it to be the superior model in both simulated data and real administrative data.

4.1.1. Background

Cost and cost savings are becoming increasingly important for the US healthcare system. During the first 5 months of 2019, over one hundred papers with a key word of “cost savings” were published in major US journals; 5 in the New England Journal of Medicine, 4 in JAMA, 3 in PLoS ONE, and the balance in other journals aimed at decision makers. This overwhelming focus on cost savings makes it imperative that we begin to standardize our approach to cost analysis, because different approaches have different underlying statistical assumptions and often result in different outcomes.

Appropriate health care cost estimation is crucial as it is used to guide evidence-based health policy implementation. Health policy makers rely on costs to drive their decisions (Power & Eisenberg). The estimation for costs associated with a disease, such as diabetes, can influence the allocation of resources for the prevention and treatment of the disease (Fukuda, Ikeda, Shiroiwa, & Fukuda, 2016). As numerous administrative data sources have become available for analysis, it is essential that these data are analyzed properly. These claims data, or billing data, are usually observational data sources, such as Medicare billing data, that are often assessed for health care cost outcomes. Inaccurate cost analysis can lead policy makers to make sub-optimal decisions.

Analysis of large datasets of observational data come with its own challenges. Data are often skewed and group comparisons often biased. Cost data are commonly positively skewed with a few outlying patients, often those with complications or who are very ill, disproportionately consisting of much of the total costs (Dodd, Bassi, Bodger, & Williamson, A comparison of multivariable regression models to analyse cost data, 2006; Bohl, Blough, Fishman, Harris, & Phelan, Are generalized additive models for location, scale, and shape an improvement on existing models for estimating skewed and heteroskedastic cost data? , 2013; Malehi, Pourmotahari, & Angali, Statistical models for the analysis of skewed healthcare cost data: a simulation study, 2015; Kurz C. , 2017). Administrative data can often include zero costs or zero observations (non-users) that may make up a large proportion of the data (Malehi, Pourmotahari, & Angali, Statistical models for the analysis of skewed healthcare cost data: a simulation study, 2015; Kurz C. , 2017). Further challenges arise when analysis of cost data also need to consider censored data within the research time frame (Dodd, Bassi, Bodger, & Williamson, A comparison of multivariable regression models to analyse cost data, 2006). Healthcare cost can vary according to region, health care system, population, and payer; thus, adjustment of costs is needed for comparisons across these factors. There are many different methods currently being used to estimate costs including: generalized linear models with a log link, natural logarithm transformed costs, gamma distribution, median regression, two-part models, and Bayesian models. As cost drives many health care policies, inaccurate analyses of cost can have serious consequences.

A review of approaches that compare cost analysis methods, which are conducted statistically and mathematically, is essential to provide needed evidence as to which methods are the most appropriate and valid for the evaluation of claims data. Therefore, we conducted such a systematic review to identify what methods of cost analysis are statistically and mathematically appropriate to use with large claims data and, specifically,

determine whether one method could be considered optimal. The knowledge gained from this review can be used to guide evidence-based cost analysis and properly assist policy makers' decisions.

4.1.2. Methods

Potential papers were identified by searching two databases, Scopus and Ovid, using the following search terms: ["methodology" or "simulations" or "bootstrap" or "bootstrapped" or "model comparison" or "compare models" or "markov" or "statistical model"] and ["cost data" or "claims data" or "billing data" or "health insurance data" or "health insurance claims" or "billing claims" or "administrative data"] and ["cost analysis" or "health care costs" or "hospital costs"] in the title, abstract, or keywords. Searches were conducted for all articles published up to the end of February 2018.

Only statistical method papers using multivariable modelling of cost, with or without methods controlling for selection bias, were considered eligible given they met one of the following inclusion criteria: 1) a comparison of two or more statistical methods to analyze cost or 2) one statistical method performed on two or more different types of cost data. Since administrative cost data can be very different from cost data collected in a randomized controlled trial, papers that used only data from randomized controlled trials and not administrative data were excluded. Additional exclusions included not being published in English and review papers. Bayesian methods and joint modelling methods were considered beyond the scope of this review. Data elements extracted from eligible papers include: paper's first author, year of publication, statistical methods being evaluated, data types used in analysis, year(s) the data was collected, simulation approach used, sample size of data, distribution of the data, the results of the method performance and the author's recommendations based on their results.

The titles and abstracts of all papers found utilizing the search terms in both databases were reviewed for potential eligibility. Among those titles and abstracts meeting

potential eligibility, the full text article was reviewed for eligibility to be included in the systematic review. The reference lists of the included papers were also assessed to identify additional eligible papers. A second reviewer verified the initial potential eligibility of papers by reviewing 20% of the full list of the titles and abstracts originally obtained from the search keywords in the Scopus database. This systematic review follows the guidelines according to Preferred Reposting Items for Systematic Reviews and Meta-Analysis (PRISMA) (Liberati, et al., 2009). A meta-analysis was not assessed due to the nature of the results of this review.

4.1.3. Results

The search of databases identified 1,048 potential papers, after title and abstract review there were 80 papers that met initial inclusion criteria for full article review (Figure 4.1.6.1). There were nine papers that were included in the review and summary data extracted (Table 4.1.6.1). A little more than half (55.5%) of these papers did not include simulations to determine the best statistical methods for modelling cost. One paper only looked at simulations and did not analyze statistical methods with real data. Of the eight papers using real cost data, all but one (87.5%) indicated a positively skewed distribution in their data.

4.1.3.1 MODELS

Of the nine papers, 28 models were assessed. The most commonly assessed models were the ordinary least squares (OLS) and generalized linear model (GLM) using the gamma distribution with log link; the latter used in comparison for all but two of the papers reviewed. All other models were predominately assessed in one paper each. Thus, comparisons of model performance between studies will be limited to OLS and GLM using gamma distribution with log link. Details of the models and findings from each study are provided below.

Dodd et al. compared multivariable models of cost; normal linear regression (NLR) of untransformed costs, NLR of log transformed costs (normal and Duan's smear retransformations), bootstrapped LR with robust standard errors, median regression, and GLM using the gamma distribution with log link (Dodd, Bassi, Bodger, & Williamson, A comparison of multivariable regression models to analyse cost data, 2006). Analysis was evaluated on administrative data and model performance was compared using root mean square error (RMSE) and mean absolute error (MAE). NLR and median regression had the worst fit and predicted negative costs, indicating that these models do not fit skewed data well. Log transformed NLR performed well on this data; however, it overestimated mean costs. The GLM Gamma with Log link was the best fit model for the highly skewed data.

Basu et al. evaluated OLS regression of untransformed costs, OLS log transformed costs, GLM Gamma with log link, and extended estimating equations (EEE) and compared goodness-of-fit using modified Hosmer-Lemeshow test, Pregibon's link test, and Pearson's test (Basu, Arondekar, & Rathouz, Scale of interest versus scale of estimation: Comparing alternative estimators for the incremental costs of a comorbidity, 2006). Assessment was conducted in administrative data of 7,428 patients that developed heart failure. Tests showed that all OLS models and GLM Gamma were non-linear, thus, the EEE model provided the best fit.

Garrido et al. compared 2-stage least squares (2SLS) of costs, 2SLS of log costs, GLM Gamma regression with log link, and full information maximum simulated likelihood (FIMSL) models (Garrido, Deb, Burgess, & Penrod, 2012). Model comparison was evaluated by the local average treatment effect (LATE) and average treatment effect on the treated (ATET). 2SLS models did not provide accurate estimates; however, the authors recommend looking at GLM Gamma regression with log link and FIMSL models

more closely as they both perform well in complex data and have similar estimates and standard errors.

Bohl et al. evaluated OLS with log transformed models, GLM Gamma regression with log links, and generalized additive models for location, scale, and shape (GAMLSS) (Bohl, Blough, Fishman, Harris, & Phelan, Are generalized additive models for location, scale, and shape an improvement on existing models for estimating skewed and heteroskedastic cost data? , 2013). This was the earliest of the included papers to perform simulations for model comparison along with assessment in real data. Simulations were conducted using Gamma and generalized inverse Gaussian distributions. OLS had the worst estimators, whereas the GLM model had the better performance overall. When the models were evaluated using Medicare advantage data the GAMLSS model was not a good fit. The GLM Gamma regression with log links performed the best.

Akbarzadeh et al. compared Poisson, negative binomial, zero-inflated Poisson, and zero-inflated binomial regressions (Akbarzadeh, Pourhoseingholi, Zayeri, & Ashtari, 2013). Evaluation of models was conducted on survey data collected on a random sample of 1,929 participants in Iran. Analysis verified assumptions that when data had large variability such as unobserved heterogeneity or a large amount of zero costs, Poisson regression was not appropriate. However, negative binomial models were better able to handle unobserved heterogeneity. Lastly, in the case of a large number of zero costs, the zero-inflated binomial performed best.

Kuwornu et al. assessed OLS models with: no cost transformation, a log cost transformation with normal retransformation, heteroscedastic retransformation, and Duan's retransformation; robust regression, GLM models with: Poisson with log link, Gamma with identity link, and Gamma with a log link (Kuwornu, et al., 2013). Models were assessed by R^2, mean absolute predication error (MAPE), and RMSE. Administrative data was used to evaluate model performance. OLS log transformed costs with normal

retransformation had the highest average R^2 reported, the robust regression had the lowest average MAPE, and the OLS with no cost transformation had the lowest average RMSE.

Malehi et al. compared OLS with log transformation models, GLM Weibull and Gamma regression models with log links, and Cox proportional hazard models in simulated data only (Malehi, Pourmotahari, & Angali, Statistical models for the analysis of skewed healthcare cost data: a simulation study, 2015). Monte Carlo simulations were conducted with varying levels of skewness using log-normal, Weibull, and Gamma distributions with Gamma distribution having the least skewed data. The Cox proportional model performed the worst across all distributions whereas the GLM Gamma was the most superior overall. When data was highly skewed, the Weibull regression model performed better than the Gamma model, though as sample size increased, the OLS model precision approached that of the GLM Weibull and Gamma models.

Fiebig et al. used micro panels to compare OLS models to fixed effects (FE) models. Models were evaluated by mean square error (MSE) and MAPE on the ability to predict future costs and out-of-sample costs (Fiebig & Johar, 2017). Monte Carlo simulations were conducted with varying correlations between the explanatory variable and unobserved time-invariant effects. When there were no correlations between the explanatory variable and unobserved time-invariant effects, the OLS model performed better. However, when there were no time-invariant effects, the FE models were the best. When there was a correlation between the explanatory variable and unobserved time-invariant effects, FE models were better able to predict post-sample costs and OLS were better able to predict out-of-sample costs. When the models were assessed in administrative data with less than 3% of observations with zero cost, FE only performed well in post-sample prediction while OLS was superior in out-of-sample predictions.

Kurtz et al. compared the Tweedie distribution model to the Tobit, Poisson, and 2-part models using the Gamma regression with log link and generalized Gamma regression with log link (Kurz C. , 2017). They considered situations of both a high correlation between group's characteristics and a low correlation. Monte Carlo simulations were conducted with varying proportions of zero costs. The simulations showed that when less than 20% of observations had zero costs, the Tweedie model was the better fit when the correlation between groups were high. However, for low correlations between groups the Tweedie performed similarly to Gamma and generalized Gamma models when the proportion of zero costs were low; however, the two-part models were better when simulations had more than 20% of observations with zero cost. When the models were compared using real administrative data with 18.1% of observations with zero cost, the Poisson and Tobit models had a worse fit compared to the Tweedie and two-part models. As suggested by the simulation data when zero costs were less than 20%, the Tweedie model had a similar model fit as the 2-part models and may be considered an alternative approach in future research.

4.1.4. Discussion

The systematic review aimed to identify what methods of cost analysis are statistically and mathematically appropriate to use with large claims data. While most papers used established methods, there were three papers that introduced new methods. Kurtz et al. found their suggested Tweedie method could be an alternative approach to analyze cost with a small number of zero costs, though, existing methods still performed superior. Garrido et al. compared only novel models to each other, further research will be needed to compare these to existing accepted methods.

The most commonly assessed models were OLS and Gamma distribution models. As seen in Table 4.1.6.2, almost all the models that made comparison with the GLM Gamma distribution with log link found this model to be the superior model. Kuwornu et al., who

compared eight different models including OLS and Gamma distributions, stated that the OLS Log model with normal retransformation had superior performance based on the highest average R^2 over 10-fold cross-validation; however, while the average R^2 was highest at 18.77, the values varied from 6.73 to 24.76 and were not consistent. Both the GLM Gamma with identity and log links had consistent R^2 values over the 10 replications with average values of 17.04 and 14.52, respectively. The same inconsistency in values can be seen with the recommendations based on lowest RMSE values for the OLS model of untransformed costs. Again, the Gamma with identity link performed just as well as the OLS model. OLS model precision was found to approach that of the GLM Gamma models as sample size increased in simulations conducted by Malehi et al.

4.1.4.1. Simulated vs. Real Data

All but one paper used real data to demonstrate the performance of their models. While simulations are a useful tool to assess the accuracy of model performance, the use of only simulated data can lead to biased results as simulations can be structured to fit a particular statistical model for superior performance. Assessment of models in real data is necessary to evaluate performance in uncontrolled environments. Some papers provided incomplete rationales for decisions regarding data, such as excluding zero costs, which may have biased the model assessment.

4.1.4.2. Limitations

To focus the scope of this review on the most common use of administrative cost data, methods assessed in randomized controlled trial data only, those that evaluated the effect of missing data, and those that assessed joint modeling methods were excluded. This focus limited the evaluation of numerous methods in various data distributions and, thus, limited the number of model comparisons.

4.1.5. Conclusions

The GLM Gamma distribution with Log link performed most consistently as the superior model in comparisons using both simulated data and real administrative data. The literature review suggests that this is the most appropriate model to use with administrative data. Some caution is suggested when dealing with heteroscedastic data or data with high proportion of zero costs (or non-users). The Tweedie distribution is an emerging new method that may be useful in future research.

4.1.6. Appendix – Tables and Figures

Figure 4.1.6.1. PRISMA flow chart.

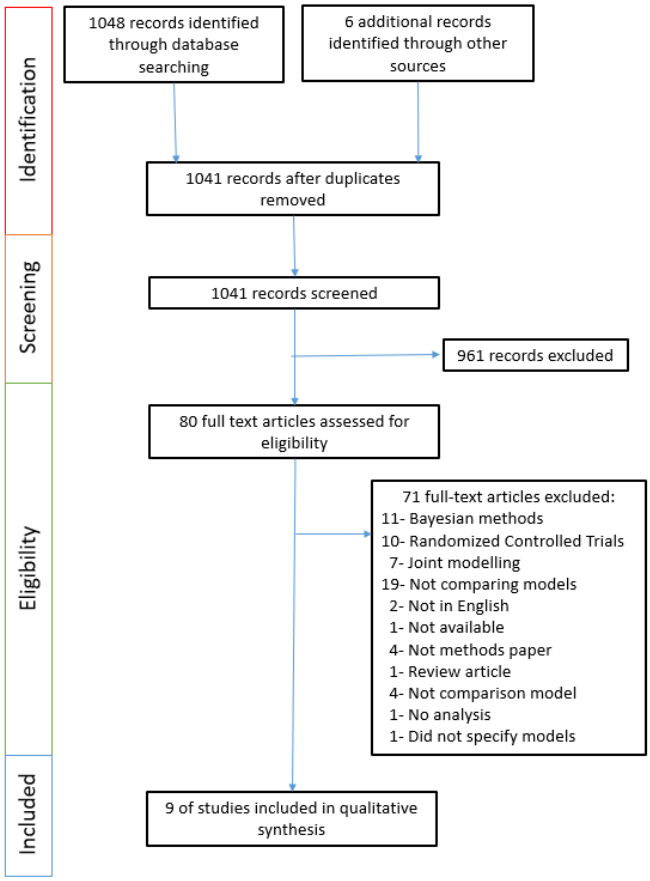


Table 4.1.6.1. Summary characteristics of methods, data, and results of reviewed papers.

First Author	Year of Publication	Methods being evaluated	Data Types	Year(s) data collected	Sample size of Data	Distribution of Data	Simulation approach used	Results of simulation	Author's recommendation
Dodd	2006	NLR NLR Log transformed Bootstrapped NLR Median regression GLM Gamma with Log link	Hospital Administrative Data	2000	N=426	Positive skewed	None	None	GLM Gamma with Log Link
Basu	2006	OLS OLS Log Transformation (D retransformation) GLM Gamma with Log Link EEE	Medstat's MarketScan	1997-2001	N=7428	Positive skewed, no zero cost	None	None	EEE
Garrido	2012	2-Stage Least Squares 2-Stage Least Squares Log Transformed GLM Gamma with Log Link Full Information Maximum Simulated Likelihood (FIMSL)	Veteran's Health Administration	2005-2006	N=3389	Positive skewed	None	None	GLM Gamma with Log Link Approach or FIMSL
Bohl	2013	OLS Log transformed (Normal retransformation) GLM Gamma with Log link GAMLSS	Medicare Advantage	Not Mentioned	N=2000	Positive skewed, no zero cost	Gamma and Generalized inverse gaussian	GLM Gamma with Log Link	GLM Gamma with Log Link
Akbarzadeh	2013	Poisson Regression Negative Binomial Zero-inflated Poisson Zero-inflated Negative Binomial	Survey	2006-2007	N=1929	Not mentioned	None	None	Zero-inflated Negative Binomial
Kuwornu	2013	OLS OLS Log transformed (Normal retransformation)	Administrative Data of Saskatchewan, Canada	1996-2010	N=17,480	Positive skewed	None	None	OLS OLS Log transformed (Normal retransformation)

		OLS Log transformed (heteroscedastic retransformation) OLS Log transformed (Duan's retransformation) Robust regression GLM Poisson with Log link GLM Gamma with Identity link GLM Gamma with Log link							Robust regression
Malehi	2015	OLS Log transformed (retransformation not mentioned) GLM Weibull with Log link GLM Gamma with Log link Cox Proportional Hazard	None	None	None	None	Monte Carlo simulated with varying levels of skewness using log-normal, Gamma, and Weibull distributions (no zeros)	GLM models performed better. Weibull performed better than Gamma with higher skewness. As sample size increased, OLS model precision approached GLM precision.	GLM Gamma was superior overall Cox was superior if Weibull distribution
Fiebig	2017	OLS Fixed Effects	Administrative Data of New South Wales	2006-09	N=264,024	Right skewed < 3% zero cost	Monte Carlo simulated with varying correlation between explanatory variable and unobserved time-invariant effects	Uncorrelated: FEP is superior except where there are no unobserved time-invariant effects Correlated: FEP is superior in Post-sample forecast	OLSP was superior

									OLSP is superior in out-of-sample forecast	
Kurtz	2017	Tweedie Tobit Poisson 2-part model: Gamma with Log link Generalized Gamma with Log link	RAND HIE	Not provided	N=3301	Right skewed 18.1% zero cost	Monte Carlo Simulated with varying proportions of non-users/zero cost	<20% zero observations: Tweedie >20% zero observations: 2-part models Gamma was better than Gen Gamma		Tweedie is an alternative to the superior 2-part models

Table 4.1.6.2. Summary Models evaluated and preferred recommendations by author of reviewed paper (denoted as **X***).

MODEL	Dodd 2006	Basu 2006	Garrido 2012	Bohl 2013	Akbarzadeh 2013	Kuwornu 2013	Malehi 2015	Fiebig 2017	Kurtz 2017
NLR	X								
Log NLR	X								
Bootstrap NLR	X								
OLS		X				**X***		**X***	
OLS Log (Heteroscedastic)		X				X			
OLS Log (Normal)				X		**X***			
OLS Log (Duan)						X			
GLM Gamma Log link	**X***	X	**X***	**X***		X	**X***		**X***
GLM Gamma Identity link						X			
GLM Generalize Gamma									**X***
EEE		**X***							
Median Regression	X								
Robust Regression						**X***			
2-Stage Least Squares			X						
2-Stage Least Squares Log			X						
FIMSL			**X***						
GAMLSS				X					
Poisson					X				X

Poisson Log Link						X			
Zero-inflated Poisson					X				
Negative Binomial					X				
Zero-inflated Negative Binomial					X				
Weibull Log							X		
Cox Proportional Hazard							X		
Tweedie									X
Tobit									X
Fixed Effects								X	
Data Type	*Real*	*Real*	*Real*	*Real/ Simulated*	*Real*	*Real*	*Simulated*	*Real/ Simulated*	*Real/ Simulated*

MANUSCRIPT 2

Minimally Important Difference in Cost Savings: Is It Possible to Identify an MID for Cost Savings?

4.2.0 Abstract

As healthcare costs continue to increase, studies assessing costs are becoming increasingly common, but researchers planning for studies that measure costs differences (savings) encounter a lack of literature or consensus among researchers on what constitutes “small” or “large” cost savings for common measures of resource use.

Other fields of research have developed approaches to solve this type of problem. Researchers measuring improvement in quality of life or clinical assessments have defined minimally important differences which are then used to define magnitudes when planning studies. Also, studies that measure cost effectiveness use benchmarks, such as cost/QALY, but do not provide benchmarks for cost differences. In a review of the literature, we found no publications identifying indicators of magnitude for costs. However, the literature describes three approaches used to identify minimally important outcome differences: 1) anchor-based, 2) distribution-based, and 3) a consensus-based Delphi methods. In this exploratory study, we used these three approaches to derive minimally important differences for two types of resource measures common in costing studies for: 1) hospital admissions (high cost); and 2) clinic visits (low cost).

We used data from two (unpublished) studies to implement the minimally important difference (MID) estimation. Because the distributional characteristics of cost measures may require substantial samples, we performed power analyses on all our estimates to

illustrate the effect that the definitions of “small” and “large” costs may be expected to have on power and sample size requirements for studies.

The anchor-based method, while logical and simple to implement, may be of limited value in cases where it is difficult to identify appropriate anchors. We observed some commonalities and differences for the distribution and consensus-based approaches, which require further examination. We recommend that in cases where acceptable anchors are not available, both the Delphi and the distribution-method of MID for costs be explored for convergence.

4.2.1. Background

Concern about the growing spending trend in healthcare (Dieleman, Cao, A, & et al, 2020) has prompted clinical and health policy decision makers to continually assess benefits and value of new treatments and care processes with an objective to control costs without sacrificing quality of care (Blumethal & Abrams, 2020; Baicker & Chandra, 2020). Rising healthcare costs are a national problem, and as part of efforts to control costs, studies to assess the effect of systems changes on costs abound (Hong, Nguyen, Yasay, & et al, 2020; Farford, Pantin, Presutti, & Ball, 2019). However, few cost studies report a formal power analysis, and the literature is silent on questions related to the magnitude of cost. This may lead to inefficient study designs with excessive sample sizes. This happens if we use a cost measure that does not provide the maximum power to detect a difference; for example, using total cost over some time period, instead of disease specific cost, or if we fail to consider known sub-group differences in expected cost as part of the randomization or analysis plan. Furthermore, cost is a complex study variable, because it may be viewed from an organizational finance or accounting perspective (fixed and variable costs and budget impact) by some decision makers, or from an economic perspective (opportunity cost or cost effectiveness) by other stake holders. Thus, it may be important to specify an MID for cost in a study in such a way that has “face validity” as

an MID of cost differences from both a financial management and an economics perspective.

We found no studies indicating how to determine “big” or “small” cost savings, even for common measures, such as hospital admissions or primary care visits. This absence of a common understanding about the magnitude of meaningful cost savings is detrimental to good planning for health care program evaluations, quality improvement assessments, and for randomized studies that assess the value (cost and consequences) of innovative health systems changes. If we do not know how big “Big” is, we may design studies that are either under powered or inefficient, neither of which is desirable. Developing a common understanding and language needed to discuss the concepts of magnitude related to cost of care is needed, because health policy research relies on statistical significance tests (p-values or confidence intervals) to judge the likelihood that cost differences, associations, and effectiveness demonstrated in our policy studies are unlikely to be due to chance. In this paper, we will 1) discuss the methods described in the literature used to define MIDs, 2) use two common resource use categories (hospital admissions and outpatient visits) that are important drivers of cost in many studies as examples for applying MID approaches to cost data, 3) show how cost MIDs behave when used in power analyses to inform study planning, and 4) make recommendations for issues to be explored further as our understanding of the usefulness of MIDs for cost improvement.

4.2.1.1. Literature of MIDs

The determination of a minimally acceptable difference for clinical measures can be easily assessed through repeated use and clinician experience from observations of the outcomes to identify what is clinically important. Other focus areas, such as quality of life improvement, use guides to judge effect magnitude. The concept of a minimal clinically important difference (MCID) was developed by Jaeschke, Singer, Guyatt to create

interpretability of the change in score of Quality of Life (QOL) questionnaires (Jaeschke, Singer, & Guyatt, 1989). A MCID, or as later referred to as minimally important difference (MID), is defined as the smallest difference perceived as beneficial that would result in a change of the patient's management (Guyatt, et al., 2002). MIDs are identified by three methods: 1) anchor-based, 2) distribution-based, and 3) consensus-based (Guyatt, et al., 2002; King, 2011).

The anchor-based method maps the relationship between the change in score of the inconclusive assessment (target) with an independent measure (anchor) that has an already established meaningfulness and an association with the target (Guyatt, et al., 2002). The anchor seeks to quantify the changes in score into trivial, small, moderate, or large categories. However, one important point of this method is that it recognizes that the same absolute difference in score may have different meaning across different portions of the scale. As an example, a 10-point change from 20 to 30 is likely to mean something different to patients and clinicians than a 10-point change from 90 to 100. Thus, according to Guyatt, interpreting results in ways that consider the proportion of patients achieving the incremental benefit may be more important than simply comparing mean differences (Guyatt, et al., 2002).

The distribution-based method of determining MID examines the relationship between the magnitude of effect and variability (Guyatt, et al., 2002). Typically, this is expressed as a ratio called Cohen's D, where the magnitude is within patient difference and variability is between patient variability for the control group at baseline, or the pooled variability of control and treatment groups at baseline (Guyatt, et al., 2002; King, 2011). There are two inherent limitations to this method to be considered. First, variability of a measure is different for each study, thus, effect sizes may not be comparable across different populations with varying levels of homogeneity (Guyatt, et al., 2002). Second, the interpretability of an effect size, in terms of a fraction of a standard deviation, may not be

easily understood by many practicing clinicians and may therefore lack face validity for clinical relevance. Cohen sought to address the latter limitation by suggesting that commonly observed study differences encompass ranges of 0.2 SD, 0.5 SD, and 0.8 SD for effects considered small, moderate, and large changes, respectively. There has been some discussion in the literature about the arbitrariness of these cut points. However, studies have provided evidence that suggest the plausibility of Cohen's ranges and consistency of standard deviations and MID within the same instrument.

The final method for specifying a MID involves a consensus or Delphi approach using expert opinion. This approach was pioneered by the Rand Corporation in the 1950s, where researchers recognized that expert judgement is often needed to solve complex problems when a definitive conclusion is not obvious (Pill, 1971; Okoli & Pawlowski, 2004; de Villiers, de Villiers, & Kent, 2005). The Delphi approach uses expert opinion refined through a series of rounds (King, 2011). The objective of the method is to distill the judgement of a panel of selected experts in a field using a process that is minimally susceptible to bias from the experts' personal characteristics, such as persuasiveness, perceived status and charisma (Okoli & Pawlowski, 2004). The Delphi method is conducted anonymously using questionnaires sent by mail, e-mail, or fax. Responses are summarized and returned to the experts for re-evaluation, until consensus is reached, or until it becomes clear that experts truly differ on this issue. Expert panel reconsiderations under conditions of anonymity reduces bias that may occur in face-to-face discussions where dominant personalities within a group of experts may sway the group's expressed opinion (Okoli & Pawlowski, 2004). The anonymity used in the Delphi method reduces bias from dominant experts while capitalizing on their knowledge and insights through repeated rounds with new responses based on summaries. The iteration provides an opportunity for group members to provide feedback and explain their choices and for

individuals to reevaluate their choice when given information provided by other experts (Okoli & Pawlowski, 2004).

The use of MIDs for studies of patient-reported outcomes (PROs) has matured (Revicki, Hayes, Cella, & J, 2008). It is recommended that MIDs be based on responses and anchors that are correlated at >.30.

4.2.1.2. Application of MIDs to Cost Data

It is expected that MIDs may vary by context, and that a single MID may not be sufficient for all study applications. Further, it is recommended that a MID should be based on multiple approaches and triangulation of methods. It is also reported that different methods for estimating MIDs often converge, and that a Delphi process is employed to select MIDs that are relevant to a study (Revicki, Hayes, Cella, & J, 2008).

While any of the three approaches for the development of MIDs are accepted as being appropriate for informing the planning of studies using QOL and clinical assessment tools, it is not known whether these approaches could be translated into use for healthcare utilization and costs. To examine how MIDs behave for cost studies, we chose to use cost differences for hospital admissions and clinic visits as examples of high-cost and low-cost study outcomes. These costs were chosen because these two types of units are often the relevant resources on which interventions to reduce cost of care are focused. We used (unpublished) data from studies of actual patient cohorts and their recorded cost measures to specify MIDs and then used these MID specifications to examine the statistical power and sample size variability imposed by each MID definition.

We examined the literature of MIDs for PRO studies to identify relevant criteria for judging how our MIDs perform. Criteria discussions often stressed measurement validity markers (construct validity, responsiveness) that are not central issues for cost studies. Ideally a MID for cost would be: 1) constant across similar types of cost “drivers”; 2) relevant for a specific costing perspective, and 3) be stable over a reasonable cost horizon.

However, the assessment of MIDs on these criteria is outside the scope of this study. So, instead of judging our MIDs against specific criteria, we used the convergent approach recommended by Revicki and colleagues (2008) combined with a pragmatic comparison on power and sample size. The literature of MIDs for PROs recommends the use criteria for choosing MIDs that: 1) are based on selection of relevant ranges that emerge when results are presented graphically; 2) weigh anchor-based results heavier that results from other methods; 3) seek convergence between methods; and 4) use a modified Delphi approach for development of consensus. The choice of describing MIDs as they behave with regard to statistical power and sample size was pragmatic based on relevance to the planning of cost studies of relevance to population health, health systems reengineering and quality improvement efforts. This focus is supported by the statement Revicki and colleagues (2008) for the use of MIDs for PRO research, that “MIDs are clearly useful for calculating statistical power and for determining sample sizes for clinical trials”.

4.2.2. Methods

The hospital admission cohort consists of patients identified as having an opioid-related event treated in any hospital in a state over a 3-year period. The clinic visit cohort consists of cost data from outpatient visits incurred over 12 months for HIV-infected adolescents from 4 clinics in different states in the US. These data were de-identified and are part of ongoing exploratory studies deemed non-human research by our IRB. The data are governed by data use agreements and not available for other use.

4.2.2.1. Anchor-based

The anchor-based MID was calculated based on the relationship of clinic care costs with the 2017 Medicare medical fees for the median (50^{th} percentile) cost for complex clinical visits. The median medical fees for visits of complex (CPT 99204) and very complex (CPT 99205) clinic visits were used, because the patient cohort utilized for the low-cost study was comprised of clinically complex patients. The complex visit is

defined as a 45-minute visit with a median payment of $293, and very complex visit is defined as a 60-minute visit with a median payment of $373, a meaningful payment difference of $80 for 15 minutes. There are no assessments in the literature of clinically meaningful cost differences in hospital admission; thus, only clinic visit cost data were assessed using the anchor-based MID method.

4.2.2.2. Distribution-based

MIDs calculated using a distribution-based method were based on Cohen's cutoffs of 0.2, 0.5, and 0.8 standard deviations (SD) for small, medium, and large effect size (ES), respectively (Cohen, 1988). Comparison ES were calculated as a percentage of mean (5%, 10%, 20%) for small, medium, and large ES, respectively. We chose 20% as the maximum change of the cost parameter, because that proportion is commonly used for sensitivity analysis in economic studies to assess value of interventions (Taylor, 2009), and our published cost effectiveness models have been shown to be robust to sensitivity analysis employing a 20% change in cost valuation across different clinical trials and country settings (Simpson, Baran, Kirback, & Dietz, 2011; Simpson, Jones, Rajagopalan, & Dietz, 2007). We chose the lower percentages for this parameter to be half and one quarter of the 20% value.

4.2.2.3. Consensus-based

MIDs calculated using a consensus-based method were based on the judgement of professionals from various backgrounds in an academic institution that assessed cost evaluations through a questionnaire that was administered via email. The 17 professionals evaluating the questionnaire included 10 faculty (2 finance, 1 health services research, 2 management, 1 policy, 1 informatics, and 1 public health); and seven practitioners (2 hospital administrators, 2 medical practice managers, 1 community health center director and 2 clinical managers). The questionnaire consisted of 4 case scenarios, two low-cost (clinic visits) and two high cost (hospital admissions) scenarios. These scenarios included

only mean and standard deviation parameters to mimic costs reported in general research papers. For the visit scenarios, the mean cost reported was $335, and the standard deviation was $237; for the two high-cost (hospital admission) scenarios, the mean cost was $18,400 and the standard deviation was $47,900. Low-cost and high-cost each had one scenario based on a sample size of 100 and one on a sample size of 1,000. For each scenario, participants were asked to rate the level of cost savings (based on examples that we derived from MID estimates $17, $35, $47, $67, $80, $120, $190 and $900, $1800, $3600, $9600, $24000, $38000; for low-cost and high-cost respectively) as one of the following effect sizes: trivial, small, medium, and large. This approach was not a true Delphi method as there were no additional rounds to form a consensus. However, the results provided narrow bands of estimates and may reflect that an underlying consensus may be reachable with few iterations.

4.2.2.4. Power Analysis

Power to detect differences were calculated for sample sizes of 100 to 1,000. Power was calculated using 1-sided independent t-tests for the anchor-based method and 2-sided independent t-tests for the distribution and expert-based methods, all at a 0.05 alpha level. All effect size and power were calculated using log (base 10) transformed costs, as the costs were positively skewed and did not meet normality assumptions. All power calculations were conducted using SAS 9.4 (Cary, NC).

4.2.3. Results

Costs for both the hospital and clinic visit studies had positively skewed distributions (Fig. 4.2.7.1). The hospital admissions study was made up of n=6,427 patients with a mean cost of $18,418 (SD = $47,908) and a median cost of $2,591. The clinic visit study was made up of 60-minute visits for n=421 patients with mean cost of $335 (SD= $237) and a median cost of $237 (Table 4.2.7.1).

4.2.3.1. Anchor-based

The anchor-based method using a meaningful payment difference of $80 between the complex 45-minute visit and the very complex 60-minute visit represents an ES of 0.34 SD and 24% of mean costs. For consistency between graphic displays across methods, Figure 4.2.7.2 shows the power calculations of the log transformed clinic visit costs for sample sizes 100 to 1,000 based on the difference of $80 under the MID anchor-method with respect to the typically minimally accepted power of 80% as indicated by the red line.

4.2.3.2. Distribution-based

Power calculations of the log transformed hospital admission and clinic visit costs, respectively, were conducted for sample sizes 100 to 1,000 based on the small, medium, and large ES under the MID distribution-based method on standard deviations and percentage of mean costs. For the hospital costs, the power to detect differences for a given sample size based on the distribution-method for MID converges to the power needed to detect differences using the percentage of the mean (Fig. 4.2.7.3) demonstrating the desired convergence across methods. This is encouraging and indicates that the use of percentages and standard deviations to define MIDs may be similar and thus could be used interchangeably or selected in a manner such that they are most relevant or where data availability drives the choice.

In this high-cost study, we found that the distribution-based method was similar to the method of using a percentage of mean costs. However, under these coefficient of variance assumptions, the sample size needed for a study to be powered to detect a difference using distribution-method of MID is the more conservative calculation when mean cost is small, as is the case with visit costs (Fig. 4.2.7.4).

4.2.3.3. Consensus-based

Consensus-based MIDs were calculated based on the 17 questionnaire responses received (85% returned). Two of the missing responses were due to clinicians working on

the front line during the COVID-19 pandemic. One missing response was a faculty member. Table 4.2.7.2 shows the average cost savings rated as small, medium, and large effect sizes under each of the four scenarios.

The mean cost values from the surveys were calculated as a proportion of the standard deviations and means from the scenarios (Table 4.2.7.3). Both the low cost and high-cost proportion of SD are much smaller proportions than the 20%, 50%, and 80% cutoffs for small, medium, and large effect size, respectively, suggested by Cohen. The average rated cost savings among low cost clinic visits as a percentage of mean and SD remained consistent between the two sample size scenarios, indicating that there was a consistent estimate of what comprises a meaningful difference for low cost studies. However, the average rated cost savings among high cost hospital admissions as a percentage of mean and SD remained consistent only among the small and medium ES between the two sample size scenarios. The disagreement is most observable in the large ES as a percentage of the mean; for the smaller sample size, a large ES was considered as much as the mean, whereas, for the larger sample size, it was almost three-quarters of the mean.

The consistency of the proportion of SD between the scenarios of n=100 and n=1000 is evident in the power calculations for both low and high cost as indicated by the closeness of the solid and dotted lines within each color (e.g. green) representing the effect size for small, medium, and large (Fig. 4.2.7.5).

4.2.4. Summary of Findings for the Three Methods

The main results for the three approaches to the specification of MID for cost studies are shown in Figure 6 below. It appears that the distribution-based method of MID is a usable approach for specifying effect sizes for cost studies. It may be superior to the use of a percent decrease in mean cost for low unit cost studies, because it would guide the researcher towards choosing a more conservative sample size estimate for measuring

cost savings. The anchor-based method mapped almost directly to the 50% SD ES of the distribution-based method (Fig. 4.2.7.6) and may prove to be useful if well-defined anchor values are available. We used the difference in payment for a CPT code increment as the anchor. However, we were not able to identify a reasonable anchor value for the hospital admission scenario. This may limit the usefulness of the anchor method for many studies.

Several important issues emerged from our study. First, as illustrated by the power curves in figure 6, studies that measure savings in low-cost resources, such as medical office visits which have large variances, are rarely adequately powered for testing hypotheses of cost differences if they have < 100 observations, unless they had lower variability than was used in our scenarios. For these types of studies, even a “large” cost difference determined by the distributional method is the only MID specification with > 80% power to detect cost savings. Large sample sizes specified by the Delphi and Anchor methods will require about 200 observations to achieve > 80% power. Studies aiming to detect “medium” sized savings specified by the Delphi method may need at least 400 observations, and small savings may require between 800 and 1,400 observations for adequate power. Figure 4.2.7.7 below shows similar patterns for MID estimates for the hospital admission cost data. Only, the very high costs and large SDs make sample size requirements much larger. Indeed, we observed that 1,000 observations are inadequate for a study where the expected cost savings are defined by a small or medium size difference, as determined by the Delphi method. Further, to achieve 80% power to find a significant difference for a “small” effect defined by the distribution method may be expected to require about 800 observations.

4.2.5. Limitations

We found that the Delphi-method MIDs were smaller than the distribution-based MIDs (Fig. 4.2.7.6 and Fig. 4.2.7.7); however, the Delphi-method used here was not extended for several rounds to achieve complete consensus among the respondents.

Thus, it is possible that the Delphi-method and distribution-based MID could have converged with further iterations. The power and sample size simulations were limited to one study for each hospital admissions and out-patient visits. The Delphi method was conducted on a convenience sample and did not include multiple rounds of the survey's results to come to a final consensus among respondents. Further simulations need to be conducted to confirm this trend.

4.2.6. Discussion and Conclusions

Our perspective in this paper is that, for health services researchers to make informed decisions about the value of new treatments or process improvements from a population health perspective, we need information on changes in outcomes and costs. Cost is a complex study variable, because it can be viewed from an organizational finance or accounting perspective (fixed and variable costs and budget impact), or from an economic perspective (opportunity cost or cost effectiveness). The value of specifying an MID for cost in a study may be that it will require us to consider cost differences from both perspectives. However, when cost is used to examine value from an economic perspective, it should be defined as a difference in the arithmetic mean cost for the populations treated (Glick & D, 2015). This cost measure may be expected to have a non-normal distribution that is skewed with a long heavy right tail, and sometimes with a large number of zero values (Glick et al, 2015 pg. 97). Standard deviations for population costs are often equal to or even greater than their mean values. Thus, compared to the distributions for health outcome measures, costs may be expected to require a much greater sample size to achieve sufficient statistical power when compared to the clinical outcomes. This poses practical as well as potentially ethical issues. From a practical perspective, an increased sample size requirement leads to increased study costs and perhaps also longer time before current practice is improved. This potentially deprives

patients of better outcomes, which may be considered unethical. Further, increasing study costs may also be considered unethical in a world of scarce resources (Williams, 1992).

In studies of program evaluation or organizational quality improvement, we may not be able to increase sample size. This means that understanding the relationship between study power and the definition of a MID specification for cost is important for study planning. Given the high likelihood that the cost measure in a study greatly affects study power, we compare our MID specification on this metric, not because it is the “best” comparator, but because it may be expected to be the greatest constraint on the study design and should be considered early in the study planning phase. In addition, the use of study power as a metric for comparing different definitions of an MID illustrates to readers the complex relationships that exist between the recommended measures of cost (arithmetic mean and SD), sample size, and a study’s power to make inferences about the value gained by the intervention (Glick et al, 2015 pg. 110). There are no firm criteria for choosing a superior method for MID. Indeed, it may be desirable to use more than one approach and examine convergence. However, three criteria should be considered: 1) the chosen method should be relevant to decision makers, 2) the method should fit the type of preliminary cost data available prior to the study (e.g. cannot use the anchor method if cost data is not available for the anchor, or if an anchor that is acceptable to the major stake holders cannot be identified), and 3) the method should reflect current benchmarks (if any) that are available for the condition in the literature.

Additional research must be conducted to determine if the Delphi-method, when fully implemented, agrees with the distribution and anchor-based method findings. Based on the limited results of this small study, it appears that the anchor-based method, while logical and simple to implement, future research should focus on identifying appropriate anchors. It may be the case that it becomes easier to define anchors as we begin to think in-depth about defining cost MIDs. Our thinking has evolved over the time of this study.

When we planned this work, we could not think of a good anchor for hospital admissions. However, now, after much discussion, we recommend that future studies of the MID for hospital cost explore the potential value of using Diagnosis-Related Group (DRG) incremental payment differences as MID anchors for hospital admissions. We therefore recommend that all three methods for defining the for MID for costs be explored further and be examined for convergence. Other issues of importance have emerged as a result of this exploratory work. From the responses to our Delphi survey, it appears that it may be important to examine if the absolute value of costs affect decision makers perception of cost savings. Health services research cost studies have a number of different audiences, and the MIDs may vary between decision-making groups. Despite the low sample size, we observed potential clustering of responses within similarly trained groups. MIDs defined as "Big" by clinicians, accountants, and administrators may differ. Thus, future studies should explicitly examine if MIDs differ by current responsibility or professional training of respondents to Delphi surveys.

4.2.7 Appendix – Tables and Figures

Table 4.2.7.1. Distribution of cost data for hospital admissions and clinic visits in USD

	N	**Mean**	**SD**	**Median (Min; Max)**
Hospital Admissions	6427	$18,418	$47,908	$2,591 (1; 1,206,879)
Clinic Visit	421	$335	$237	$237 (50; 1,617)

Table 4.2.7.2. Results from Survey of Decision Makers in the Consensus-based Approach: Mean (SD) Value for Survey Responses Classifying Cost Savings by Effect Size given a Specified Sample Size (USD)

	Low-Cost		**High-Cost**	
Sample Size	*100*	*1000*	*100*	*1000*
Effect Size				
Small	$35 (13)	$43 (26)	$1,920 (750)	$1,575 (402)
Medium	$68 (25)	$68 (31)	$5,453 (3,242)	$4,388 (2,681)

Large	$131 (55)	$120 (59)	$18,141 (11,989)	$13,418 (10,977)

Table 4.2.7.3. Average Survey Cost Savings as proportion of standard deviation and mean

	Low-Cost[a]				**High-Cost[b]**			
N	*100*		*1000*		*100*		*1000*	
	SD	Mean	SD	Mean	SD	Mean	SD	Mean
Effect Size								
Small	15%	10%	18%	13%	4%	10%	3%	9%
Medium	29%	20%	29%	20%	11%	30%	9%	24%
Large	55%	39%	51%	36%	38%	99%	28%	73%

[a] Mean = $335, SD = $237

[b] Mean = $18,400, SD = $47,900

Fig. 4.2.7.1. Histogram of hospital costs (left) and clinic visit costs (right) in US dollars

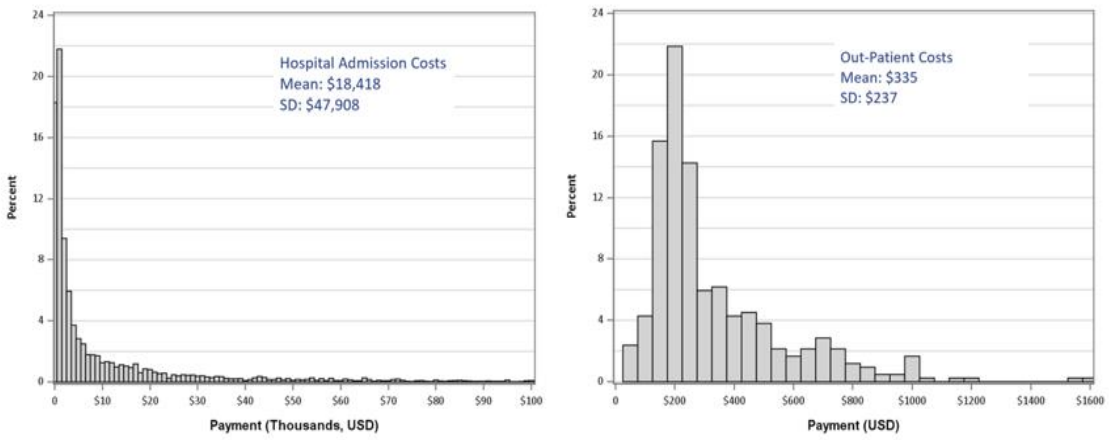


Fig. 4.2.7.2. Power calculations of log transformed hospital costs for sample sizes 100-1,000 based on percentage of standard deviation and mean with 80% power level

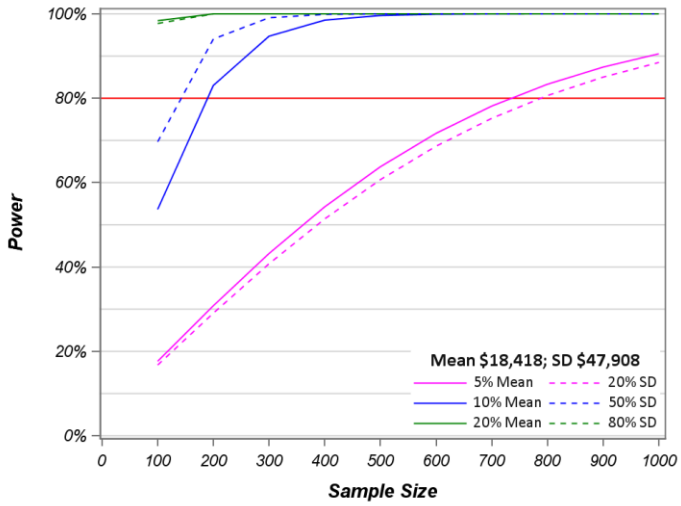


Fig. 4.2.7.3. Power calculations of log transformed visit costs for sample sizes 100-1,000 based on Cohen's d suggested cutoffs and percentage of mean with 80% power level

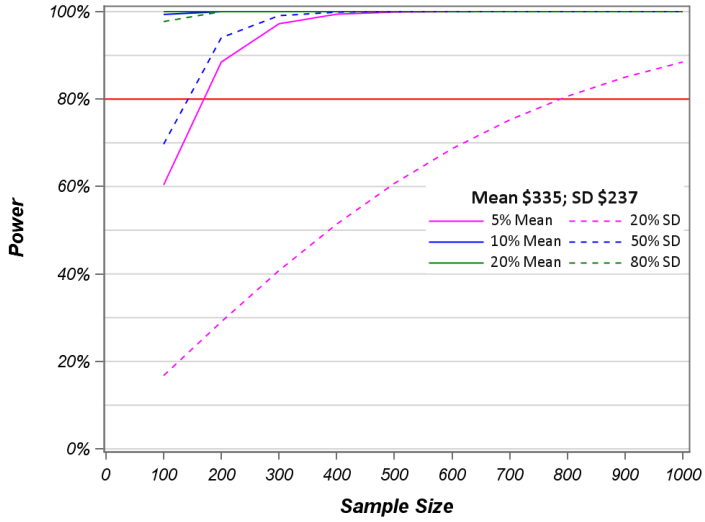


Fig. 4.2.7.4. Power calculations of log transformed costs for sample sizes 100-1,000 based on rated cost savings as a percentage of SD for low-cost clinic visits (left) and high-cost hospital admissions (right) with 80% power level

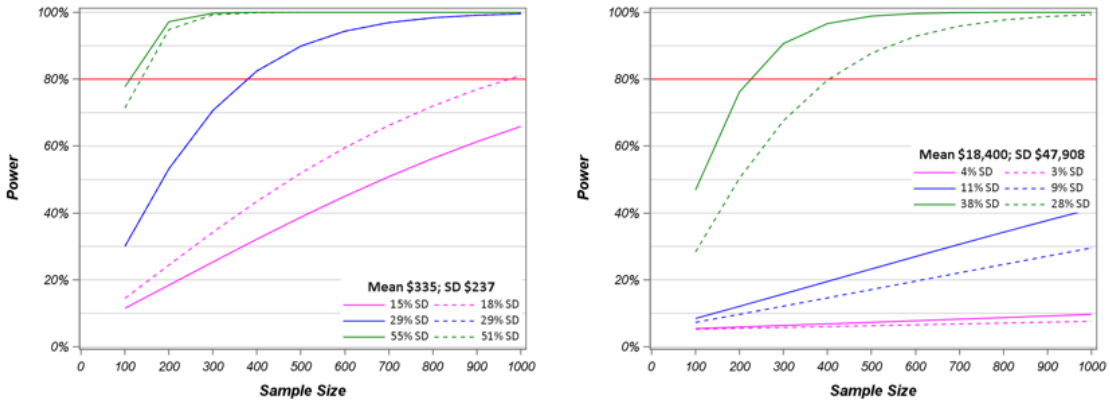


Fig. 4.2.7.5. Power calculations of log transformed costs for sample sizes 100-1,000 based on rated cost savings as a percentage of SD for low-cost clinic visits (left) and high-cost hospital admissions (right) with 80% power level

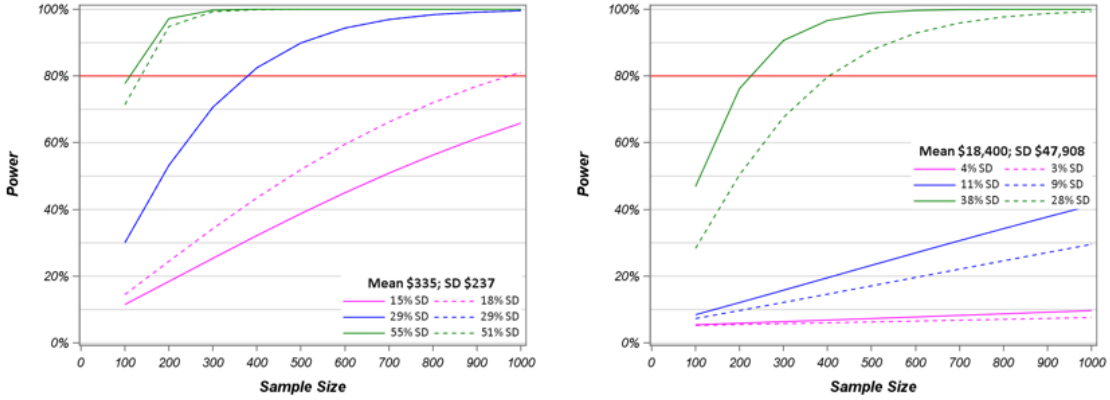


Fig. 4.2.7.6. Power calculations of log transformed visit costs for sample sizes 100-1,000 based on the anchor, distribution, and Delphi methods with 80% power level

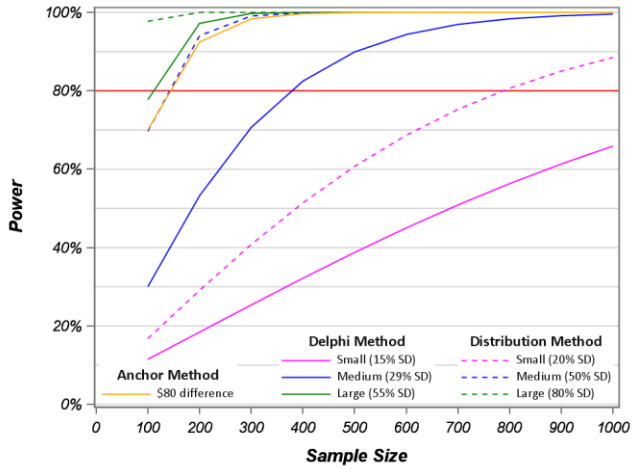
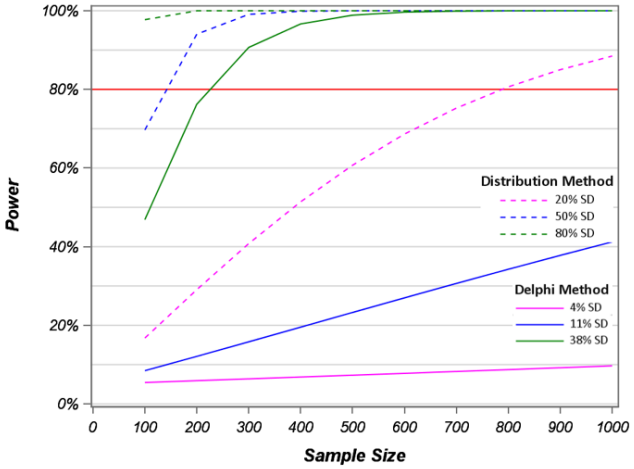


Fig. 4.2.7.7. Power calculations of log transformed hospital admission costs for sample sizes 100-1,000 based on the distribution and Delphi methods with 80% power level



MANUSCRIPT 3

Comparing cost of telehealth and in-person visits using time-drive activity-based costing (TDABC)

4.3.0 Abstract

Background: Cost studies of telehealth and virtual visits (TH) are few and report mixed results of the economic impact of virtual care and telehealth. Largely missing from the literature are studies that identify the cost of delivering TH versus in-person care.

Materials and Methods: The objective is to compare cost of delivering virtual and in-person care for a pediatric sick-visit example using a modified time-driven activity-based costing (TDABC) approach. We examined visits before and during COVID-19 using: 1) recorded structured interviews with providers, 2) iterative workflow mapping; 3) EHR time stamps for validation; 4) standard cost weights for wages; and 5) clinic CPT billing code mix for complexity weighs. We examined the variability in estimated time using a decision tree model and Monte Carlo simulations.

Results: Workflow charts were created for the clinic before COVID-19 and during COVID-19. Using TDABC and simulations for varying time, the weighted cost of clinic labor for a sick visit before COVID-19 was $54.47 versus $51.55 during COVID-19.

Conclusions: Our TDABC approach is feasible to use under virtual working conditions; requires minimal provider time for execution; and generates detailed cost estimates that have “face validity” with providers and are relevant for economic evaluation.

4.3.1. Background

When the COVID-19 pandemic struck, telehealth and virtual visits (TH) became essential for both patients and providers. The urgent need to convert in-person care to TH

meant few health systems had time to plan and be deliberate in their TH approach. TH programs already in place were scaled up quickly and improvisations were common. We are now at the stage where we must make strategic decisions for a streamlined, sustainable TH approach and identify the best opportunities for improvement. The cost and value of TH services developed during the pandemic scale-up should inform our choices. Unfortunately, cost/economic studies of TH are few and report mixed results of the economic impact of virtual care and telehealth. Reports include large variations in prices (Nguyen, 2020), increased overall cost of care due to “convenience” effect of virtual visits (Jain & Mehrotra, 2020), and cost savings accrued from reduced travel time, improved triage and access to timely care (Hooshmand & Foronda, 2018). These published studies report costs from the perspective of patients, payers, and populations. Largely missing from the literature are studies that identify the cost of delivering telehealth versus in-person care from a provider perspective.

As healthcare costs increase at alarming rates, there is a need to have accurate information when making decisions based on the value (cost-effectiveness) of health interventions and health care processes. Decision makers must determine the most efficient allocation of limited resources while delivering the best quality of care. Typically cost analysis is evaluated using a top-down approach and may not be representative of the true costs of care (Carroll & Lord, 2016; Zilberberg & Shorr, 2010). Nonetheless, these costing results often guide the decision-making about health care process changes. Better costing methods are therefore needed to provide a more accurate true cost estimate to make better informed decisions. Time-driven activity-based costing (TDABC) is a less frequently used micro-costing methodology that more accurately identifies costs of production using service specific activity and resource use evaluated with patient specific treatment times (Gammon & Cotten, 2016; Carroll & Lord, 2016; Tan, Rutten, van Ineveld, Redekop, & Roijen, 2009). The TDABC method has been utilized to identify areas for

process improvement, though, it has not always been cost efficient to perform (Shander, et al., 2010). The biggest barrier to using traditional TDABC is that it is resource intensive; requiring research staff with expertise in what to assess to be present in the clinic to observe and record the care processes of each patient using a stopwatch to manually collect the timing of processes and resource use.

The objective of this project was to compare the weighted labor costs of an in-person clinic sick visit before COVID-19 to the in-person and telehealth clinic sick visit during the COVID-19 pandemic using a modified TDABC approach in a pediatric clinic.

4.3.2. Materials and Methods

We assessed cost of providing an in-person vs. mix of telehealth and in-person sick visits for at a pediatric clinic in a suburban area. The study met institutional definition of a quality improvement project and did not require IRB oversight. A mixed methods approach was used for data collection and analysis to perform a modified TDABC of a sick visit that could be assessed in a virtual environment. A sick visit was defined as a low complexity clinic visit (CPT 99213), classified as a 15-minute face-to-face visit. Visits were described for children between the ages of 5-9 years old. The TDABC steps included: 1) recorded structured interviews with providers, 2) iterative workflow mapping, 3) EHR timestamps for time validation, 4) standard cost weights for wages, 5) clinic CPT billing code mix for complexity weights and 6) simulations to assess effects of uncertainty on cost differences.

4.3.2.1. Interview data

Structured interviews were used to collect data to map the care process of a sick clinic visit for an established patient age 5-9. An interview guide with five questions with probes was used (Appendix A). Interviews were conducted by two interviewers familiar with the workflow and EHR system. Two providers were interviewed separately: a physician (MD) and a nurse practitioner (NP). Both interviewers were present for both

sessions and the interviews were recorded. The recorded interviews were processed using Rapid Qualitative Analysis (Taylor, Henshall, Kenyon, Litchfield, & Greenfield, 2018) to develop the workflow charts. Iterative review of recordings was used to reach agreement on clinic flow between interviewers. The resulting flow charts were reviewed and edited as needed by providers who then gave final approval of their position-relevant charts.

4.3.2.2. EHR data

Two independent sources of data from the clinic were used to extract time stamps for clinic visits with CPT 99213. Mean (SD) in minute values were used to validate the minute estimates in the clinic flow charts and to identify visits with CPT billing codes for tests and time stamps for validating prescription-related process effects. These data were used to estimate the mix of visits to generate complexity weights for the cost estimates. One set of process validation data was extracted from the EPIC telehealth dashboard, used by practice managers and telehealth personnel to monitor the processes in the clinic. Data were extracted for all clinic pediatric patients with a low complexity clinic visit (CPT 99213) seen during September 2020. These data included the timestamps for check-in, treatment start time, provider treatment team composition (i.e. MD or NP), record access by each actor, and printing timestamp used by providers to present a visit summary and care plan for patients at the end of the visit. A second data set was extracted from the EPIC Clarity Warehouse, which included all CPT 99213 clinic visits in September 2019 and September 2020. These data were used to validate virtual visit time stamps and to estimate in-person visit time stamps to be used for estimating visit costs.

4.3.2.3. Workflow Mapping Conventions Applied

The clinic flow mapping process for patients utilizing the clinic for a sick visit, from signing-in at the beginning of the appointment to the conclusion of the clinic visit, began with the review of the recorded interviews. Identification of each step in the process was

completed along with the determination of actors (e.g. MD, NP, nurse, front desk personnel) and approximate time in minutes to complete each step. Three workflow charts were then developed for 1) in-person clinic visits before COVID-19, 2) telehealth clinic visits during COVID-19, and 3) in-person clinic visits during COVID-19. For each of the workflow charts, steps of the process (identified by a square) are organized by the order in which they are completed. Potential additional or alternative steps are identified with decision nodes (diamond) in the flow chart. For each step of the process, the average time to complete is noted (contained within small circle in bottom right corner of squares), and actors involved (color-coded) are identified and listed. The three workflow charts were created from each interview, then the interviews were reviewed again to make edits to the workflow charts. The two interviewers then met to review the workflow charts and both interviewers agreed on the construction of the charts.

The workflow chart was reviewed while listening to the recorded interviews and checked to be sure there were no missing connections and then returned to the interviewees for verification of accuracy. Once any suggested edits were completed, the interviewers reviewed the recorded interviews to ensure the accuracy of the workflow charts. This process was repeated as analysis of minutes and cost were conducted. The workflow charts generated by the MD and NP were combined to create a single clinic process workflow chart for each visit type before and during COVID-19. To validate the workflow process and time estimates for TDABC, EHR access during the visit were used as a proxy for contact time of actors with the patient. EHR data from the EPIC dashboard were used to evaluate the timestamp of record access and actors (Appendix B). This is the step that replaces in-clinic time estimation in traditional TDABC and makes this a modified TDABC approach.

4.3.2.4. Workflow Cost Calculation

Using TDABC costing methods, the labor cost of care was estimated for in-person clinic visits before COVID-19 and the telehealth and in-person clinic visits during COVID-19. Labor costs were calculated using median US salaries for each actor using the US Bureau of Labor Statistics salary data for 2019. Total loaded salary for each actor was calculated as median salary plus fringe benefits, equal to 35% of the median salary. A total of 2,080 annual hours worked were assumed for a full-time employee; a position-specific work capacity rate was applied: 1) nursing (e.g. Licensed Practical Nurse (LPN) and Certified Medical Assistant (CMA)) and administration staff at 80% (1664 hours) and 2) for provider (e.g. pediatrician and nurse practitioner) at 72.3% (1504 hours). The cost per minute for each actor was calculated as the total loaded salary divided by the number of capacity hours per year divided by 60 (minutes). For process steps that were either completed by two different actors or may potentially be completed by two different actors, as identified in the workflow chart, a 50/50 weight was given to each actor's salary to estimate the cost per minute for the time in the mixed process step.

The cost of each actor for the visit is determined by the total minutes utilized multiplied by the cost per minute for the actor. For in-person and TH sick visit, the labor costs across all actors are summed to determine the total labor cost of the clinic visit. Analysis of labor cost was conducted using Microsoft Excel.

4.3.2.5. Simulations

Monte Carlo simulations were developed to mimic the variation of labor minutes by the providers and staff, thus the variation of total labor cost, in the clinic setting. The provider time was simulated using a Weibull distribution identified by Medicare specified range for the visit type (Centers for Medicare and Medicaid Services, 2020). All other staff time was modeled on a Beta PERT distribution defined by a 'most likely' value (estimated

time from the flow chart) and a minimum and maximum value (±10% of the estimated time). Median salary for actors were varied on a normal distribution with 10% of the median salary for the standard deviation. The variation in minutes and cost per minute for actors provided a distribution of costs across 100,000 visits for in-person visits before COVID and telehealth and in-person visits during COVID. Simulations were conducted using Crystal Ball software.

4.3.2.6. Final Costing and Decision Tree

EHR data extracted from EPIC Clarity warehouse were used to identify and categorize providers to determine a provider for the same week in September for 2019 and 2020. Providers for each of the clinic visits were categorized as MD or NP. Decision trees were constructed using the identified provider mix and delivery method mix (in-person or telehealth visit) to calculate an average weighted visit cost using the mean and +/- 1 standard deviation estimated from the simulations.

4.3.2.7. Cost Comparison

Mean and standard deviations of forecasted visit cost estimated by simulations were evaluated for differences between before COVID and during COVID. Minimally important difference (MID) measured by a well-defined anchor has been identified as a conservative effect size for low-cost studies (Dooley, Simpson, Nietert, Williams Jr., & Simpson, 2021). The anchor-based MID was based on the relationship of clinic care costs between the low complexity sick visit (CPT 99213), defined as a 15-minute face-to-face clinic visit, and the moderate complexity sick visit (CPT 99214), defined as a 25-minute face-to-face clinic visit. The median Medicare medical fee in 2017 was $125 and $184, for low and moderate complexity visits, respectively. We selected the CPT anchored MID as a meaningful payment difference between the two adjacent visit CPT codes which is $59.

4.3.3. Results

4.3.3.1. Workflow Chart

Workflow charts capturing visit resource use estimates before COVID-19 were similar in process and time. The combined workflow used an average of estimated times for each step where times differed. Additionally, the workflow charts for each provider of a telehealth visit during COVID-19 were almost identical. However, the workflow created for an in-person clinic visit during COVID-19 had variations in the process. The risk of a patient with a suspected COVID-19 infection for the in-person clinic sick visits during the COVID-19 phase had the greatest variation which depended on need for use of PPE. We chose to use the most conservative process for time estimates (i.e. assuming donning and doffing PPE for each encounter instead of remaining in the same PPE all day).

Before COVID-19, the estimated time a provider was involved in the clinic visit between treatment start time and visit summary printing was 15 minutes, however, the addition of a laboratory test (e.g. nasal or throat swab) or prescription ordering each added an additional minute to the clinic process (fig. 4.3.6.1). The average time for an LPN or CMA was four minutes, however, if a laboratory test is needed then an additional 10 minutes is utilized by either an LPN or provider for specimen collection and analysis. The total labor time varied between 19 to 31 minutes.

Telehealth clinic visits during COVID-19 involved only providers. The estimated total labor minutes were 18, unless the provider encountered internet link issues with the patient that required them to verify information and assist patient with the connection. However, some telehealth visits are aborted. If it is determined during the telehealth visit that the patient needs to be seen in-person then the telehealth visit is canceled so that the patient does not get billed twice. However, labor costs must be counted for these visits and allocated across all telehealth visits. This is because the provider must conduct the telehealth assessment, schedule an in-person visit, and provide notes for the in-person

clinic provider that will be seeing the patient. This unbilled visit takes an estimated 21 minutes of the provider's time (fig. 4.3.6.2).

The clinic process for an in-person visit during COVID-19 for a non-COVID risk patient is almost identical to the process before COVID-19. However, the in-person clinic visit with a COVID risk resulted in less overall labor time due to fewer actors involved. These visits are completed by the provider alone to reduce staff COVID exposure. For a sick visit that does not require ordering laboratory tests or prescriptions, the estimated labor time for the process is 18 to 19 minutes, regardless of the method of delivery or the effect of the pandemic (fig. 4.3.6.3).

4.3.3.2. Visit Cost Calculation

Total loaded salary and cost per minute for each actor were calculated from median US salaries (Table 4.3.6.1). Using time-driven activity-based costing methods, the labor cost for a sick clinic visit was calculated using established workflow charts (Table 4.3.6.2). The labor costs for in-person visits before and during the pandemic were similar, however, the process became more efficient during the pandemic and results in a slightly lower labor cost ($56.16 vs. $54.68 for MD and $38.23 vs. $31.63 for NP, before COVID and during COVID pandemic respectively). Though TH visits were mostly restricted to only provider time, the provider spent almost the same time on the visit but the elimination of additional actors resulted in lower labor costs ($54.68 to $49.61 for MD and $36.75 to $31.63 for NP, under in-person and telehealth visits respectively).

4.3.3.3. Simulations

We used Monte Carlo simulations executed in Excel with a Crystal Ball extension specifying 100,000 estimates to examine the effect of potential variations in minute and costs on the estimates. Table 4.3.6.3 reports the mean and standard deviation from the Monte Carlo simulations developed with the distribution the labor costs fit for each of the 3 workflows assessed under the MD and NP teams. For clinic visit costs before COVID,

provider minutes varied 14.7 minutes and higher with labor costs fitting Gamma distributions for both MD and NP visits. The distribution from the simulations for visit labor costs for MD ranged from $36-$80 (Fig. 4.3.6.4). The clinic visit costs for Telehealth during COVID both fit Beta PERT distributions with a lower cost range for MD visit costs than in-person visits (fig. 4.3.6.5). The estimated in-person clinic visit costs fit a Beta PERT distribution for MD visits and a Gamma distribution for NP visits. The distribution for the MD in-person visit costs had a slightly narrower range during COVID than before COVID (fig. 4.3.6.6).

4.3.3.4. Final Costing and Decision Tree

The provider mix at the clinic was identified from EHR as 83.2% MD and 16.8% NP for visits. Using this provider mix with the estimated costs, the mean weighted cost per visit at the clinic before COVID was $54.47 with a range in costs of $47.26-$61.68 within 1 SD (Table 4.3.6.4). During COVID, the delivery method mix identified from EHR of 28.3% telehealth and 71.7% in-person was added to the decision tree to provide an overall weighted labor cost across both provider mix and delivery mix (fig. 4.3.6.7). The mean weighted cost per visit at the clinic during COVID was $51.55 with a range in costs of $44.38-$58.73 within 1 SD.

4.3.3.5. Cost Comparison

The difference in mean weighted visit costs were well below the MID of $59 that we had specified as important for this study. A limitation of our analysis was the inability to account for the number of TH visits that were aborted because the provider decided that the problem warranted an in-person visit. An increase in this rate would increase the weighted labor cost during COVID. To examine the effect of this factor we performed a sensitivity analysis. Assuming an 80/20 mix of telehealth and aborted telehealth to in-person in the decision tree, the mean weighted visit cost during COVID would be $54.47, the same mean weighted visit cost as observed before COVID. However, if we assume

the addition of a 20% aborted telehealth visit to in-person visit to the decision tree resulted in the identified 28.3/71.7 delivery method mix seen, the true delivery method would actually have a rate of 35.5/64.5 with a mean weighted visit cost during COVID being $54.84.

4.3.4. Discussion

This study has demonstrated that using our modification of TDABC the estimated mean labor cost for care during the pandemic has not changed from the pre-COVID period. This lack of change is largely due to the increased use of TH, which reduced provider time by allowing them to perform several tasks simultaneously, such as chart review, during their virtual encounters instead of before entering the exam room. However, our results indicate that TH may be underutilized, and that provider organizations should look to find an optimal mix of TH and in-person visits. Further, the mix of MD and NP providers should be examined to assure that the optimal mix is present for the clinic's patient severity mix. Our TDABC approach helps inform important discussions of 1) which TH programs to maintain; 2) how best to improve TH efficiency; and 3) which factors in a clinic's workflow can be changed to achieve the most efficient mix of TH and in-person visits. Simulations can be utilized to illustrate the effect of uncertainty in estimates derived for a modified TDABC approach. Its role is important to show the level of uncertainty associated with specific attempts to maximize workflow efficiency. However, it is important to use the correct distributional assumptions for the simulations or the results may be misleading. We used the assumptions embedded in the CMS specifications of expected ranges of minutes for a CPT code. Our simulation results replicate what CMS specifies to be resource ranges that are acceptable for reimbursement of provider time. Thus, this approach to choosing distributions for simulations could be useful if applied to other visit types.

4.3.5. Conclusion

As healthcare providers plan for sustained TH operations, our modified TDABC approach may be helpful. It is feasible to use under virtual working conditions, requires minimal provider time, can be implemented quickly, captures important variations in the process of care that affect costs, and generates detailed cost estimates that have “face validity” with providers and are relevant for process improvement and economic evaluation. This approach also addresses the significant barrier to the practical use of TDABC by avoiding direct observation, and replacing observed time with EHR time stamp analysis, which reduces resources required to complete the analysis and makes it implementable for processes that take place in our current “virtual” environment. While no cost difference was found in this study during the pandemic, on the other side of cost allocation that is beyond the scope of this research, is the ever evolving telehealth reimbursement, which may also impact choice of telehealth or in-person visits.

4.3.6. Appendix – Tables and Figures

Figure 4.3.6.1. Workflow chart for in-person clinic visits before COVID-19

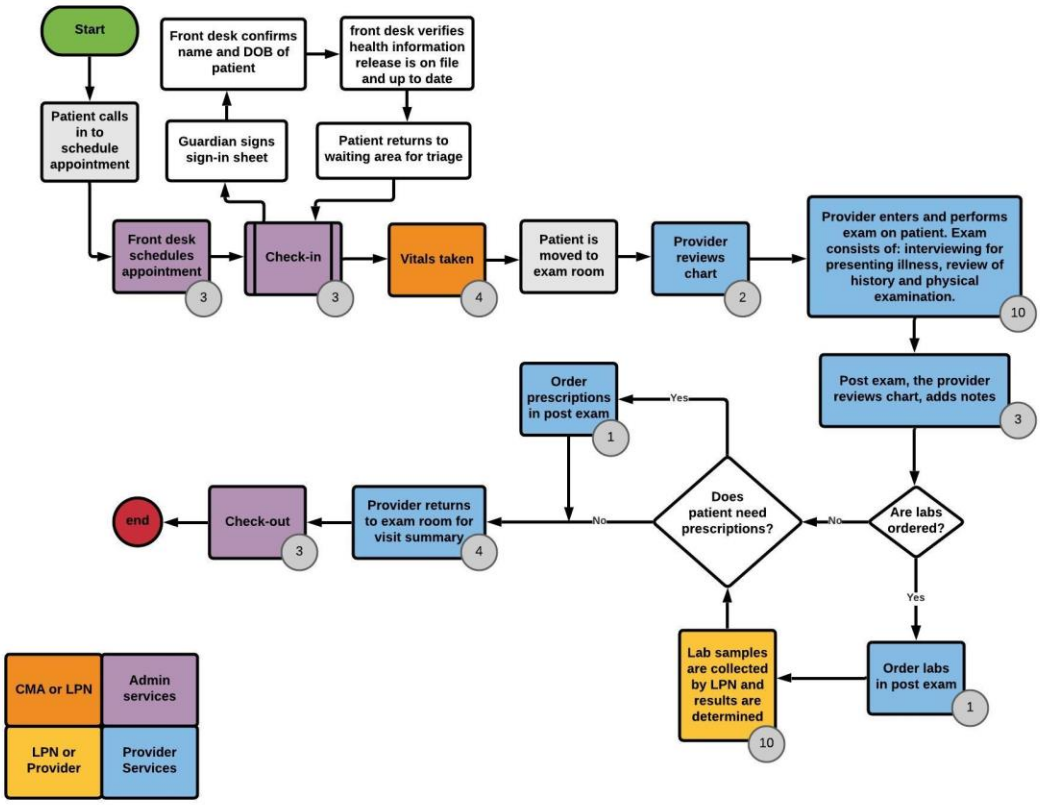


Figure 4.3.6.2. Workflow chart for Telehealth clinic visits during COVID-19

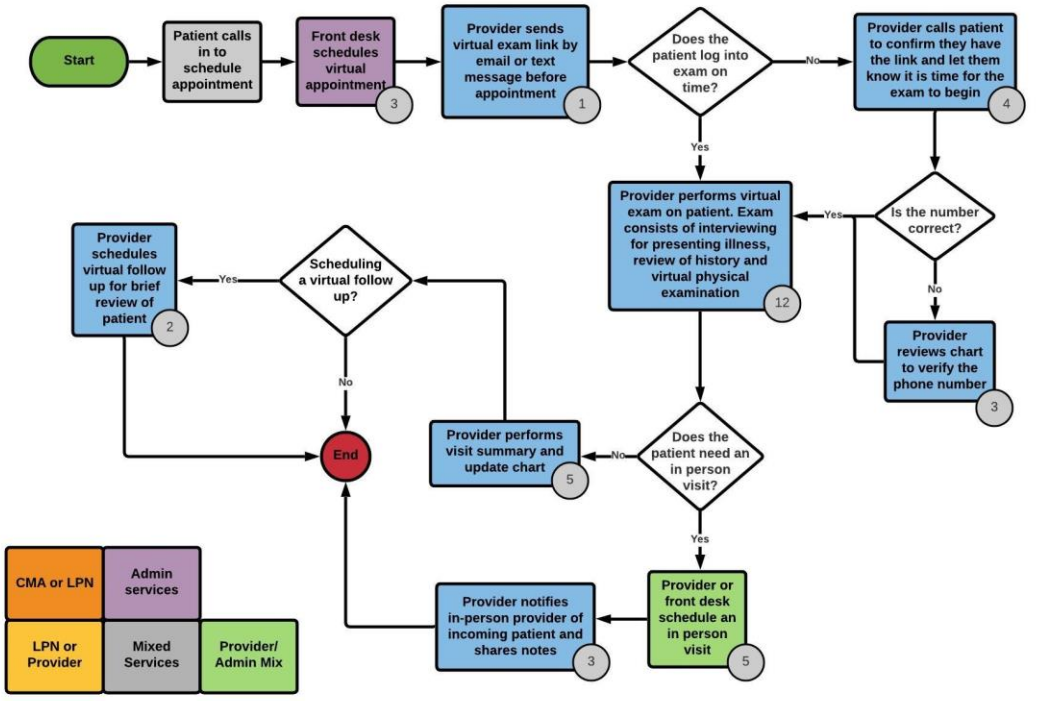


Figure 4.3.6.3. Workflow chart for in-person clinic visits during COVID-19

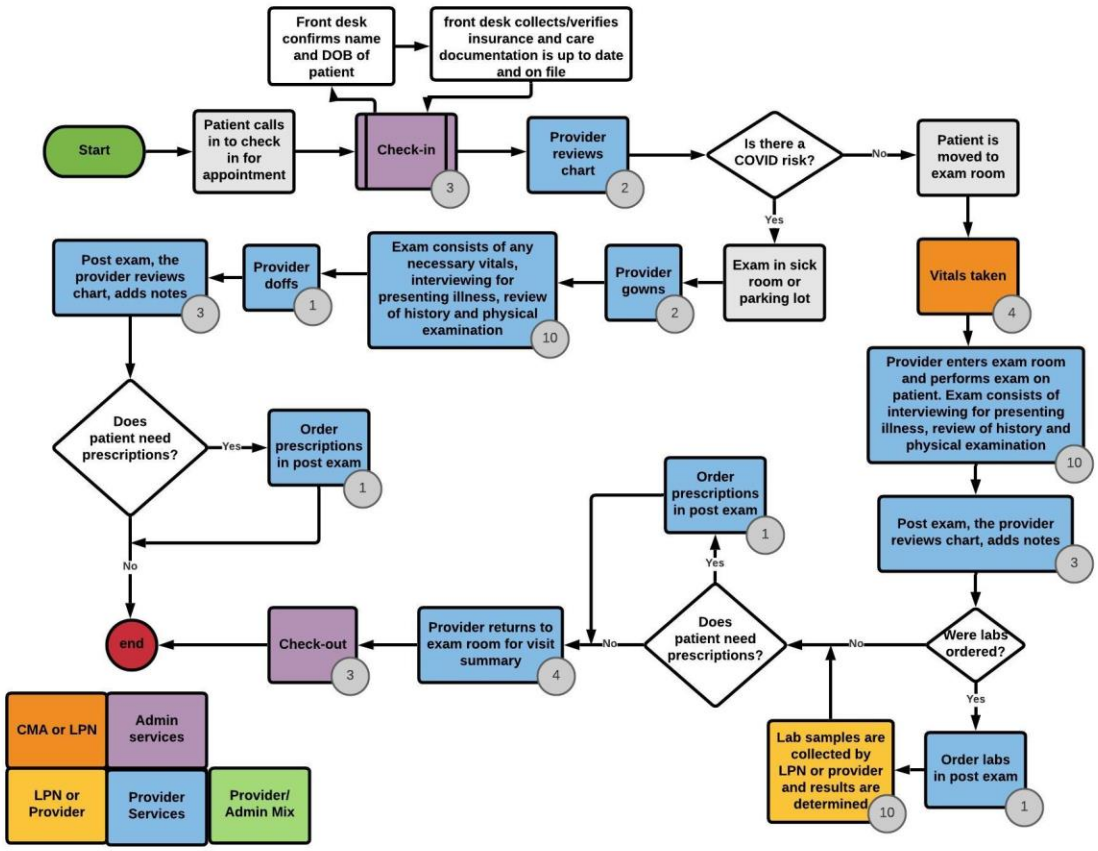


Figure 4.3.6.4. Distribution of Monte Carlo simulation for MD in-person clinic visit costs before COVID-19

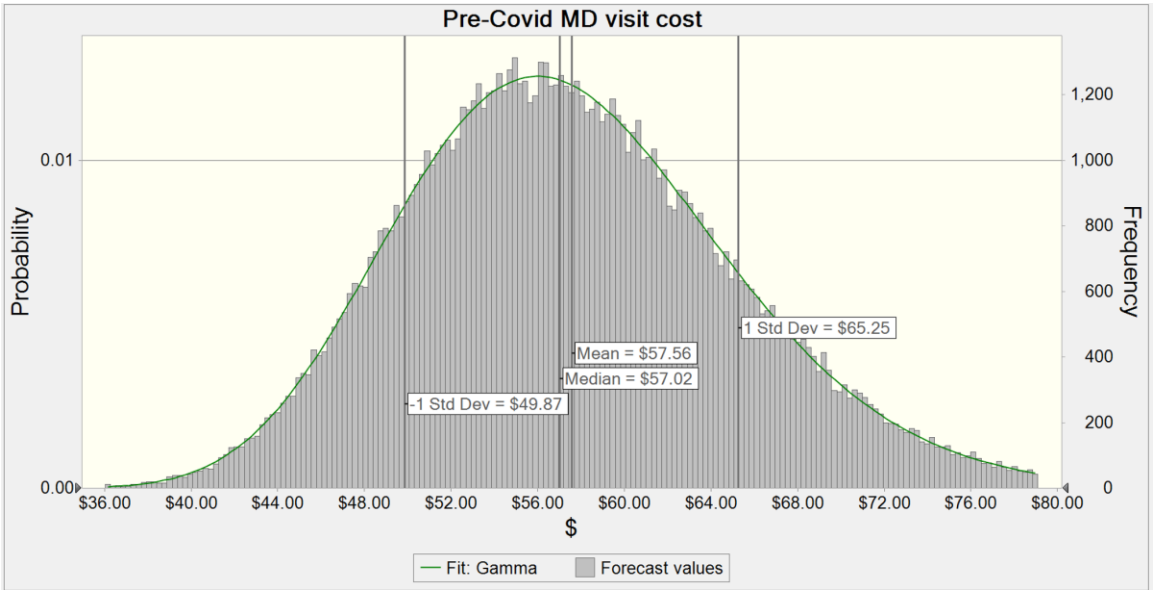


Figure 4.3.6.5. Distribution of Monte Carlo simulation for MD telehealth clinic visit costs during COVID-19

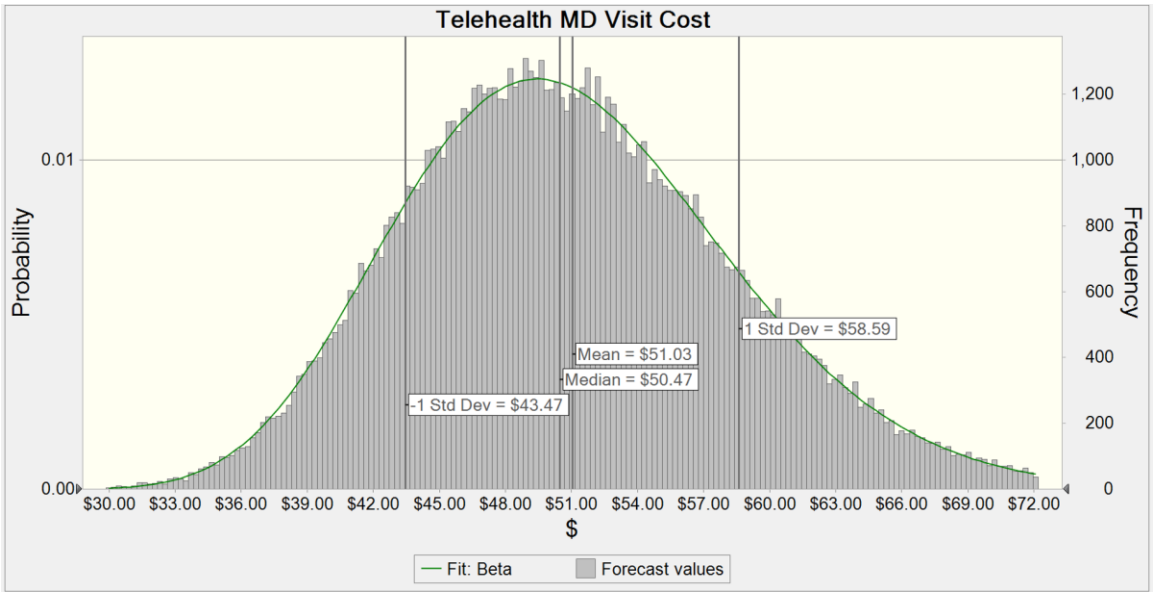


Figure 4.3.6.6. Distribution of Monte Carlo simulation for MD in-person clinic visit costs during COVID-19

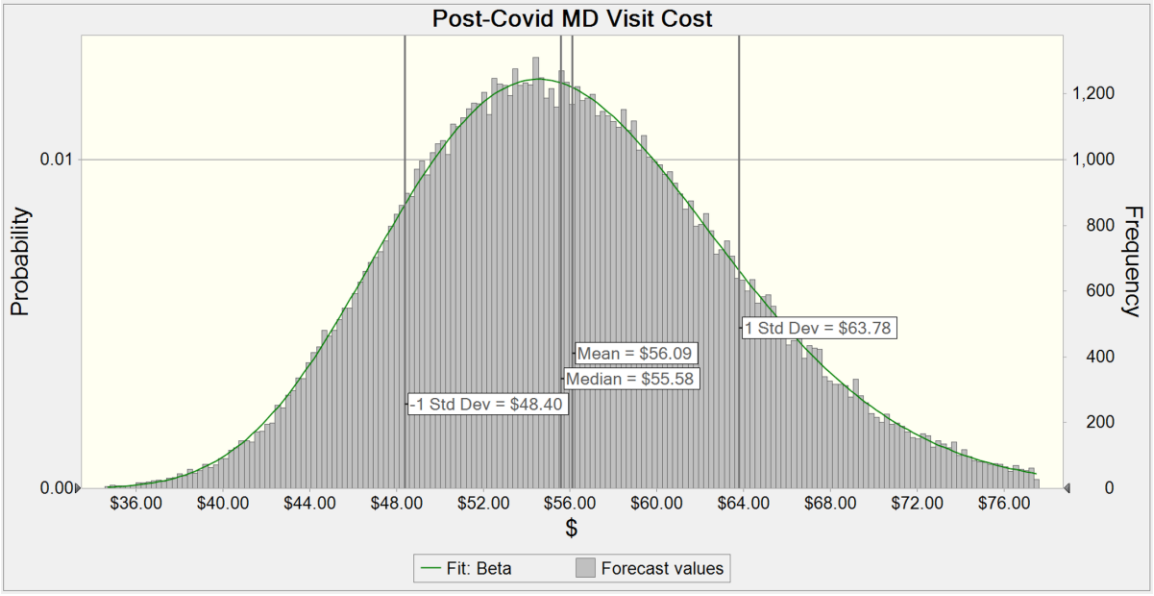


Figure 4.3.6.7. Decision Tree for in-person clinic visits during COVID-19

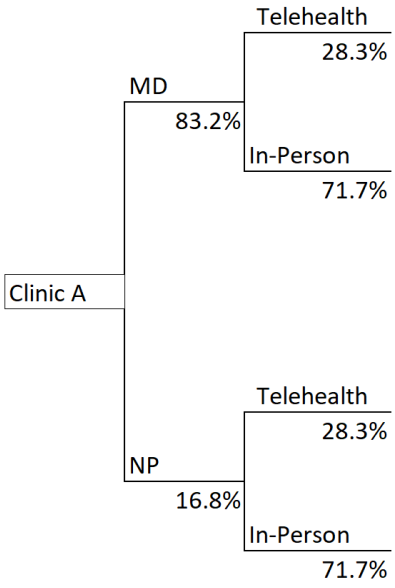


Table 4.3.6.1. Median salary costs and Cost/Minute (USD) for actors in clinic care process

Role	Median Salary	Fringe	Loaded Salary	Annual Hours	Cost/Min
Pediatric MD	$175,300	$61,355	$236,655	1,504	$2.62
Pediatric NP	$109,800	$38,430	$148,230	1,504	$1.64
CMA[1]	$34,800	$12,180	$46,980	1,664	$0.47
LPN[2]	$47,500	$16,625	$64,125	1,664	$0.64
Medical Office Assistant	$36,600	$12,810	$49,410	1,664	$0.49

[1] Certified Medical Assistant

[2] Licensed Practical Nurse

Table 4.3.6.2. Labor costs estimated for clinic visits from workflow chart (USD)

Visit type/Provider	Total Staff Min	Total Provider Min	Labor Cost
Before COVID-19			
In-person with MD	13	19	$56.16
In-person with NP	13	19	$38.23
During COVID-19			
Telehealth with MD	3	18.35[1]	$49.61
Telehealth with NP	3	18.35[1]	$31.63
In-person with MD	10	19	$54.68
In-person with NP	10	19	$36.75

[1] Link and Telephone issues assessed at 5% of visits

Table 4.3.6.3. Labor costs estimated from simulation of 100,000 visits (USD)

	Mean	SD	Fit Distribution
Before COVID-19			
In-person with MD	57.56	7.69	Gamma
In-person with NP	39.10	4.83	Gamma
During COVID-19			
Telehealth with MD	51.03	7.56	Beta
Telehealth with NP	32.49	4.72	Beta
In-person with MD	56.09	7.69	Beta
In-person with NP	37.60	4.82	Gamma

Table 4.3.6.4. Weighted labor costs forecast by Monte Carlo simulations (USD)

	1 SD below Mean	**Mean**	**1 SD above Mean**
Before COVID-19	47.26	54.47	61.68
During COVID-19	44.38	51.55	58.73
Difference	*2.88*	*2.92*	*2.95*

CHAPTER 5

DISCUSSION

5.1. Conclusion

The science of costing is well developed among economists but has not circulated to normal health services research and is especially lacking for clinical investigators beginning to look at areas such as costs. As healthcare reimbursement continues a shift to the value proposition, it is essential we enhance cost allocation accuracy; thus, we should no longer consider costs and research resources in healthcare as an “other” variable or an “other” measure. It is necessary to integrate the use of correct methods to analyze costs using the large databases, specifying clinical trials assessing economic outcomes to ensure they have an appropriate minimally important difference, and we must progress toward using micro-costing to find both cost-effectiveness and examining process innovations, such as, telehealth and use of EHR. This study showed all three areas are essential, under described in the literature, and, even if understood within a narrow group of economists that work in this area, the methods have not been disseminated well.

5.2 Future Research

Additional research must be conducted for MID in costs to determine if the Delphi-method, when fully implemented, agrees with the distribution and anchor-based method findings. Based on the limited results of this paper, it appears future costing research should focus on identifying appropriate anchors. It may be the case that it becomes easier to define anchors as we begin to think in-depth about defining cost MIDs. We therefore recommend that all three methods for defining the for MID for costs be explored further and examined for convergence. Future research that utilize simulations to illustrate the

effect of uncertainty of estimates from larger sample sizes and variations for low-cost and high-cost scenarios will assist in identifying any convergence of the MID methods in costs.

Other issues of importance have emerged as a result of this exploratory work. From the responses to our Delphi survey, it appears that it may be important to examine if the absolute value of costs affect decision makers perception of cost savings. Health services research cost studies have a number of different audiences, and the MIDs may vary between decision-making groups. Despite the low sample size, we observed potential clustering of responses within similarly trained groups. MIDs defined as “Big” by clinicians, accountants, and administrators may differ. Thus, future studies should explicitly examine if MIDs differ by current responsibility or professional training of respondents to Delphi surveys.

The modified TDABC approach is feasible to use under virtual working conditions; requires minimal provider time; can be implemented quickly; captures important variations in the process of care that affect costs; and generates detailed cost estimates that have “face validity” with providers and are relevant for process improvement and economic evaluation. Future research that can examine the true proportion of aborted TH visits that were aborted because the provider decided that the problem warranted an in-person visit could provide a more accurate weighted clinic visit cost. This approach addresses the significant barrier to practical use of TDABC by resources required to complete the analysis, thus, implementing the approach across multiple clinics can provide a better assessment of an estimated delivery method mix that would optimize labor resources and reimbursement while providing quality care. In addition, this approach to choosing distributions for simulations could be useful if applied to other visit types.

REFERENCES

Akbarzadeh, A., Pourhoseingholi, A., Zayeri, F., & Ashtari, S. (2013). Zero inflated statistical count models for analyzing the costs imposed by GERD and dyspepsia. *Arab Journal of Gastroenterology, 14*, 165-168.

Baicker, K., & Chandra, A. (2020). Do we spend too much on health care? *NEJM, 383*(7), 605-608.

Basu, A., Arondekar, B., & Rathouz, P. (2006). Scale of interest versus scale of estimation: Comparing alternative estimators for the incremental costs of a comorbidity. *Health Economics, 15*, 1091-1107.

Basu, A., Arondekar, B., & Rathouz, P. (2006). Scale of interest versus scale of estimation: Comparing alternative estimators for the incremental costs of a comorbidity. *Health Economics, 15*, 1091-1107.

Benjamin, E., Blaha, M., Chiuve, S., Cushman, M., Das, S., Deo, R., . . . Mozaffaria. (2017). *Heart Disease and Stroke Statistics- 2017 Update.* American Heart Association.

Blumethal, D., & Abrams, M. (2020). The Affordable Care Act at 10 years - Payment and delivery system reforms. *NEJM, 382*(11), 1057-1063.

Bohl, A., Blough, D., Fishman, P., Harris, J., & Phelan, E. (2013). Are generalized additive models for location, scale, and shape an improvement on existing models for estimating skewed and heteroskedastic cost data? . *Health Services Outcomes Res Method, 13*, 18-38.

Bohl, A., Blough, D., Fishman, P., Harris, J., & Phelan, E. (2013). Are generalized additive models for location, scale, and shape an improvement on existing models for estimating skewed and heteroskedastic cost data? . *Health Services Outcomes Res Method, 13*, 18-38.

Carroll, N., & Lord, J. (2016). The growing importance of cost accounting for hospitals. *Journal of Health Care Finance*, 172-185.

Centers for Medicare and Medicaid Services. (2020, November 23). *CMS.gov.* Retrieved from Centers for Medicare and Medicaid Services: https://www.cms.gov/medicaremedicare-fee-service-paymentphysicianfeeschedpfs-federal-regulation-notices/cms-1734-f

Chen, Y., Xie, W., Gunter, C., Liebovitz, D., Mehrotra, S., Zhang, H., & Malin, B. (2015). Inferring clinical workflow efficiency via electronic medical record utilization. *AMIA Annual Symposium Proceedings/AMIA symposium*, 417-425.

Claxton, G., Rae, M., Levitt, L., & Cox, C. (2018, May 8). *How have healthcare prices grown in the U.S. over time?* Retrieved from Peaterson-Kaiser Health SYstem Tracker: https://www.healthsystemtracker.org/chart-collection/how-have-healthcare-prices-grown-in-the-u-s-over-time/#item-start

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Scences, 2nd edition.* Hillsdale: Lawrence Erlbaum Associates.

de Villiers, M., de Villiers, P., & Kent, A. (2005). The Delphi technique in health sciences education research. *Medical Teacher, 27*(7), 639-643.

Dieleman, J., Cao, J., A, C., & et al. (2020). US health care spending by payer and health condition. *JAMA, 323*(9), 863-884.

Dodd, S., Bassi, A., Bodger, K., & Williamson, P. (2006). A comparison of multivariable regression models to analyse cost data. *Journal of Evaluation in Clinical Practice, 12*(1), 76-86.

Dodd, S., Bassi, A., Bodger, K., & Williamson, P. (2006). A comparison of multivariable regression models to analyse cost data. *Journal of Evaluation in Clinical Practice, 12*(1), 76-86.

Dooley, M., Simpson, A. N., Nietert, P. J., Williams Jr., D., & Simpson, K. N. (2021). Minimally important difference in cost savings: Is it possible to identify an MID for cost savings? *Health Services and Outcomes Research Methodology, 21*, 131-144.

Dudley, R., Harrell, F., Smith, L., Mark, D., Califf, R., Pryor, D., . . . Hlatky, M. (1993). Comparison of analytic models for estimating the effect of clinical factors on the cost of coronary artery bypass graft surgery. *Journal of Clinical Epidemiology, 43*(3), 261-271.

Farford, B., Pantin, S., Presutti, J., & Ball, C. (2019). Evaluation of a Family Medicine transitional care service line. *J Am Board Fam Med, 32*(4), 619-627.

Fiebig, D., & Johar, M. (2017). Forecasting with micro panels: the case of health care costs. *Journal of Forecasting, 36*, 1-15.

French, K., Albright, H., Frenzel, J., Incalcaterra, J., Rubio, A., Jones, J., & Feeley, T. (2013). Measuring the value of process improvement initiatives in a preoperative assessment center using time-driven activity-based costing. *Healthcare, 1*, 136-142.

Fukuda, H., Ikeda, S., Shiroiwa, T., & Fukuda, T. (2016). The effects of diagnostic definitions in claims data on healthcare cost estimates: evidence from a large-scale panel data analysis of diabetes care in Japan. *PharmacoEconimics, 34*, 1005-1014.

Fukuda, H., Ikeda, S., Shiroiwa, T., & Fukuda, T. (2016). The effects of diagnostic definitions in claims data on healthcare cost estimates: evidience from a large-scale panel data analysis of diabetes care in Japan. *PharmacoEconomics, 34*, 1005-1014.

Gammon, E., & Cotten, A. (2016). The efficacy of activity based accounting techniques for target case management in outpatient settings: a case study in predicting

financial risk to a nonprofit community health service provider prompted by public policy change. *Journal of Health Care Finance*, 2-12.

Garrido, M., Deb, P., Burgess, J., & Penrod, J. (2012). Choosing models for health care costs analyses: Issues of nonlinearity and endogeneity. *Health Services Research, 47*(6), 2377-2397.

Glick, H., & D, J. (2015). *Economic Evaluation in Clinica Trials 2nd ed.* Oxford UK: Oxford University Press.

Guyatt, G., Osoba, D., Wu, A., Wyrwich, K., Norman, G., & group, C. S. (2002). Methods to explain the clinical significance of health status measures. *Mayo Clinic Proceedings, 77*(4), 371-383.

Gyatt, G., Osoba, D., Wu, A., Wyrwich, K., Norman, G., & group, C. S. (2002). Methods to explain the clinical significance of health status measures. *Mayo Clinic Proceedings, 77*(4), 371-383.

Hong, Y.-R., Nguyen, O., Yasay, S., & et al. (2020). Early performance of Hospital Value-Based Purchasing program in Medicare. *Med Care, 58*, 734-743.

Hooshmand, M., & Foronda, C. (2018). Comparison of telemedicine to traditional face-to-face care for children with special needs: A quasiexperimental study. *Telemedicine & eHaealth, 24*(9), 433-441.

Jaeschke, R., Singer, J., & Guyatt, G. (1989). Measurement of health status: Ascertaining the minimal clinically important difference. *Controll Clinical Trials, 10*, 407-415.

Jain, T., & Mehrotra, A. (2020). A comparionson of direct-to-consumer telemedicine visits with primary care visits. *JAMA Netw Open, 3*(12), e2028392.

Kaiser Family Foundation. (2011, April 12). *Henry J Kaiser Family Foundation*. Retrieved from Snapshots: Health Care Spending in the United States & Selected OECD

Countries: https://www.kff.org/health-costs/issue-brief/snapshots-health-care-spending-in-the-united-states-selected-oecd-countries/

Kaplan, R., & Witkowski, M. (2014). Better accounting transforms health care delivery. *Accounting Horizons, 28*(2), 365-383.

Karp, L., Freeman, R., Simpson, K., & Simpson, A. (2018). Changes in efficiency and quality of nursing EHR documentation after implementation of an admission patient history essential dataset. *Computers, Infomatics, Nursing*.

Khatri, P., Abruzzo, T., Yeatts, S., Nichols, C., Broderick, J., & Tomsick, T. (2009). Good clinical outcome after ischemic stroke with successful revascularization is time-dependent. *Neurology*, 73: 1066-1072.

King, M. (2011). A point of minimal important difference (MID): a critique of terminology and methods. *Pharmacoeconomics Outcomes Research, 11*(2), 171-184.

Kurz, C. (2017). Tweedie distributions for fitting semicontinuous health care utilization cost data. *Medical Research Methodology, 17*(171), 1-8.

Kurz, C. (2017). Tweedie distributions for fitting semicontinuous health care utilization cost data . *Medical Research Methodology, 17*(171), 1-8.

Kuwornu, J., Lix, L., Quail, J., Wang, E., Osman, M., & Teare, G. (2013). A comparison of statistical model for analyzing episodes-of-care costs for chronic obstructive pulmonary disease exacerbations. *Health Services Outcomes Res Method, 13*, 203-218.

Liberati, A., Altman, D., Tetzlaff, J., Mulrow, C., Gotzsche, P., Ioannidis, J., . . . Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS Med, 6*(7).

Malehi, A., Pourmotahari, F., & Angali, K. (2015). Statistical models for the analysis of skewed healthcare cost data: a simulation study. *Health Economics Review, 5*(11), 1-16.

Malehi, A., Pourmotahari, F., & Angali, K. (2015). Statistical models for the analysis of skewed healthcare cost data: a simulation study. *Health Economics Review, 5*(11), 1-16.

Mandell, D., Guevara, J., Rostain, A., & Hadley, T. (2003). Medical expenditures among children with psychiatric disorders in a medicaid population. *Psychiatric Services, 54*(4), 465-467.

Manning, W., & Mullahy, J. (2001). Estimating log models: to transform or not to transform. *Journal of Health Econiomics, 20*, 461-494.

Mans, R., Schonenberg, H., Leonardi, G., Panzarasa, S., Cavallini, A., Quaglini, S., & van der Aalst, W. (2008). Process mining techniques: an application to stroke care. *Studies in Health Technology and Informatics*, 573-578.

Mazighi, M., Chaudhry, S., Ribo, M., Khatri, P., Skoloudik, D., Mokin, M., . . . Amarenco, P. (2013). Impact of onset-to-reperfusion time on strok mortality: A collaborative pooled analysis. *Circulation*, 1980-1985.

McGlothlin, A., & R, L. (2014). Minimal clinically important difference: defining what really matters to patients. *JAMA, 312*(13), 1342-1343.

Nguyen, A. (2020, March 26). *How much doesa telemedicine visit cost? A price comparison chart*. Retrieved February 25, 2021, from Good Rx: https://www.goodrx.com/blog/telemedicine-true-cost-and-telehealth-price-comparison-chart/

Nichols, G., Bell, T., Pedula, K., & O'Keeffe-Rosetti, M. (2010). Medical care costs among patients with establised cardiovascular disease. *American Journal of Managed Care, 16*(3), e86-e93.

Okoli, C., & Pawlowski, S. (2004). The Delphi method as a research tool: an example, design considerations and applications. *Information & Mangement, 42*, 15-29.

Oxman, A., Fretheim, A., Lavis, J., & Lewin, S. (2009). SUPPORT Tools for evidence-informed health Policymaking (STP) 12: Finding and using research evidence about resource use and costs. *Health Research Policy and Systems*, 7.

Pill, J. (1971). The Delphi method: substance, context, a critique and an annotated bibliography. *Socio-Economic Planning Sciences*, 57-71.

Power, E., & Eisenberg, J. (1998). Are we ready to use cost-effectivesness analysis in health care decision-making? A health services research challenge for clinicians, patients, health care systems, and public policy. *Medical Care, 36*(5), MS10-MS17.

Power, E., & Eisenberg, J. (n.d.). Are we ready to use cost-effectiveness analysis in health care decision making? A health services research challenge for clinicians, patients, health care stytems, and public policy.

Revicki, D., Hayes, R., Cella, D., & J, S. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epi, 61*, 102-109.

Sanders, G., Neumann, P., Basu, A., Brock, D., Feeny, D., Krahn, M., . . . Ganiats, T. (2016). Recommendations for conduct, methodological practices, and reporting of cost-effectiveness analyses: Second panel on cost-effectiveness in health and medicine. *JAMA*, 316 (10): 1093-1103.

Saver, J., Goyal, M., van der Lugt, A., Menon, B., Majoie, C., Dippel, D., . . . Fran. (2016). Time to treatment with endovascular thrombectomy and outcomes from ischemic stroke: a meta-analysis. *JAMA*, 316 (12): 1279-1288.

Shander, A., Hofmann, A., Ozawa, S., Theusinger, O., Gombotz, H., & Spahn, D. (2010). Activity-based costs of blood tranfusions in surgical patients at four hospitals. *Transfusion, 50*, 753-765.

Simpson, K., Baran, R., Kirback, S., & Dietz, B. (2011). Economics of switching to second-line antiretroviral therapy with lopinavir/ritonavir in Africa: Estimated based on DART trial results for Kenya and Uganda. *Value Health, 18*, 1048-1054.

Simpson, K., Jones, W., Rajagopalan, R., & Dietz, B. (2007). Cost-effectivesness of lopinavir/ritonavir tablets compared with atazanavir plus ritonavir in antiretroviral-experienced patients in the UK, France, Italy and Spain. *Clin Drug Invest, 27*(12), 807-817.

Tan, S., Rutten, F., van Ineveld, B., Redekop, W., & Roijen, L. (2009). Comparing methodologies for the costs estimation of hospital services. *European Journal of Health Economics*, 10: 39-45.

Taylor, B., Henshall, C., Kenyon, S., Litchfield, I., & Greenfield, S. (2018). Can rapid approaches to qualitative analysis deliver timely, valid findings to clinical leaders? A mixed methods study comparing rapid and thematic analysis. *BMJ Open, 8*(10), e019993.

Taylor, M. (2009, April). *http://www.bandolier.org.uk*. (Bandolier) Retrieved August 31, 2020, from http://www.bandolier.org.uk/painres/download/whatis/What_is_sens_analy.pdf

Wang, G., Zhang, Z., Ayala, C., Dunet, D., Fang, J., & George, M. (2014). Costs of hospitalization for stroke patients aged 18-64 years in the United States. *Journal of Stroke and Cerebrovascular Diseases, 23*(5), 861-868.

Weinstein, M., Siegel, J., Gold, M., Kamlet, M., & Russell, L. (1996). Recommendations of the Panel on Cost-Effectiveness in Health Medicine. *JAMA*, 276 (15): 1253-1258.

Williams, A. (1992). Cost-effectiveness analysis: is it ethical. *J Med Ethics, 18*, 7-11.

Willink, A., Reed, N., & Lin, F. (2019). Cost-benefit analysis of hearing care services: What is it worth to Medicare? *JAGS*, 1-6.

Wu, D., Smart, N., Ciemins, E., Lanham, H., Lindberg, C., & Zheng, K. (2017). Using EHR audit trial logs to analyze clinical workflow: A case study from community-based ambulatory clinics. *AMIA annual Symposium Proceedings*, (pp. 1820-1827).

Zhang, H., Mehotra, S., Liebovitz, D., Gunter, C., & Malin, B. (2013). Mining deviations from patient care pathways via electronic medical record system audits. *ACM Transactions on Management Information Systems*, 4(4): 17.

Zilberberg, M., & Shorr, A. (2010). Understanding cost-effectiveness. *Clinical Microbiology and Infection*, 16: 1707-1712.

APPENDICES

Appendix A - Interview Guide

First, we are going to talk about a visit for a sick child, think about a child between the ages of 5-9 that happened in the fall before we started any video visits.

Think about the process that would have normally taken place in your practice when that child comes in.

1. How is that child's visit scheduled?
    a. How much time does it take to schedule?
2. Now the mother and child show up at the front desk, what happens?
    a. Front desk clerical staff does that?
    b. How much time do you think it takes staff to do that?

    ** Not worried about time patients sit and wait, only staff*
3. Now you have an exam room free in the back, what happens?
    a. What measurements are taken?
    b. How much time does that take?
    c. Who does the triage? *E.g. CMA, nurse?*
4. Now the patient is sitting in the examining room, what happens?
    a. As the provider, what do you do?
        i. Before you go in, do you look at a record?
            1. How much time does that take?
    b. Now you're in the exam room, there's the mother and child, what do you do?
        i. How much time does that take?
5. You are done examining the child, what happens?
    a. Do you leave?
        i. What do you do?

        ii. How much time does that take?

    b. Do you go back into the room?

        i. If so, how long does that take?

6. Then what happens?

    a. End of visit?

        i. Who walks them out?

        ii. Check out?

            1. If so, who checks them out? Clerical staff?

            2. How long does that take?

7. What if the child needs a test? *E.g. flu swab or strep swab*

    a. How does that change the visit?

    b. Who does the swab? E.g. Nurse?

        i. If someone else, do you talk to them?

        ii. How long does the swab take?

8. The test results come back, then what happens?

    a. Positive

        i. Do you go back in?

        ii. How long does that take?

    b. Negative

        i. Do you go back in?

        ii. How long does that take?

9. End of visit the same?

    a. If no, what happens?

Now we have the same sick visit happening under COVID.

Think about the process that normally takes place in your practice for a virtual visit.

10. How is the child's visit scheduled?

- *a.* How much time does it take to schedule?

11. As the Provider, how do you know they are scheduled?

12. Is there anything you need to do before the virtual visit?

- *a.* How long does that take?

13. What do you do if the patient is late logging onto their visit?

- *a.* What do you work on in that time?
- *b.* What if you have a wrong number?
  - *i.* How often do you think that happens?

14. Now you have mom on video with her child sitting next to her, what do you do?

- *a.* How much time does the virtual exam take?
- *b.* Do you have more often follow-ups or check-ins with virtual than when in-person?
  - *i.* How do you do a check-in?
    1. How long does that take?
  - *ii.* How do you do a follow-up?
    1. How long does that take?

15. Suppose you feel the child needs a test, e.g. flu swab or strep swab, or needs to be seen in-person; what happens?

- *a.* In-person a different provider?
  - *i.* If yes, does in-person need to reexamine?
    1. Do you send notes to other provider?

16. How do you end the virtual visit?

17. Do you have anything else you need to do after the call ends?

- *a.* How much time does that take?

When a child needs to be seen in-person following a virtual visit,

What does the process look like?

18. Does the child check-in?

a. How long does that take?

19. When the nurse gets the child, what happens?

a. Are the nurses masked?

b. Put on PPE?

i. If yes, how long does that take?

We will turn this into 2-3 flow charts.

We will send these to you by email,

If they are fine and we haven't made any mistakes or left anything out, then you can just ok them by email. If there are changes, we would like to have another brief talk with you so you can explain what we need to edit.

Appendix B - Workflow Validation Using EHR Time Stamps

B.1 Methods

To validate the workflow process and time estimates for TDABC, EHR time stamps for access during the visit were used as a proxy for estimating contact time of actors with the patient. EHR data from the EPIC dashboard were extracted for one week and visually examined for relevance in use for estimating contact minutes for the visit flow charts. We conferred with clinical users and informatics experts on which timestamps would be expected to most consistently capture the visit workflow and contact times of actors. The major problem that we identified was related to the fact that no check-out time was present in the virtual visit record. When the record indicated a change in actor, but the previous actor did not have a log out time, the next actor's recorded record access was used as the end point. If the calculated time difference between record access was less than 1 minute, 1 minute of actor time was assigned. After consultations with informatics and clinical experts, we decided to use the printing of patient advisement summary timestamp used by the last step in the virtual visit. However, it should be noted that our finding of this lack of a formal "check out" has resulted in the consideration of a change in the clinical process. On examination it was found that the lack of a check out step in many cases resulted in a failure to schedule needed follow up visits. The selection of these timestamps for our study meant that we could not validate the minute estimates involved for the steps completed by the medical office assistants at check-in and check-out. This is a limitation of our study, but a minor one because the medical assistant time contribution to the care workflow is recognized to be very short and consume a low-cost resource. The error may be expected to be greater if we assumed that the medical office assistant's time was used when in actual practice the time is more likely to reflect a patient waiting for treatment to begin. This is especially the case for those patients who check-in early.

Clinic workflow actors were identified and categorized into groups with similar functional and cost characteristics (e.g. MD, NP, LPN, CMA). Other actors whose time statms were observed in the records, but who had not been identified as central actors in the workflow specification were excluded from the not typically involved in the standard clinic workflow were excluded from our time stamp analysis (e.g. human services). Visits that did not have a printing time stamp were excluded from our minute validation because the clinic visit end time could not be established. Given the positively skewed data for time and costs, mean, median and interquartile range are reported. Analysis of the time-stamp data was conducted using SAS version 9.4.

B.2 Results

There were 706 clinic visits available for analysis from the EHR data extract. The overall mean clinic visit time was 22.8 minutes (SD=19.2) with an interquartile range of 10-31 minutes. This was consistent with the estimate from the validated workflow charts. The data showed that most visits were completed by 1 to 2 unique actors (44.1% and 28.8% respectively). As the number of unique actors involved in the visit increased, the mean time of the visit also increased.

Access times across actor categories was positively skewed so medians are reported here. Median access time for MD's was accurate according to workflow chart estimates of 15 minutes. Residents had a slightly lower median access time, 11 minutes, and nurse practitioners had the lowest median access time at 9 minutes. Median access time for LPN was 14 minutes, which is accurate for clinic visits including lab tests. Whereas CMA median access time was approximately 10 minutes compared to the estimated 4 minutes in the workflow chart. This discrepancy of slightly higher LPN and CMA time may be due to not having a record log out time and result from us having to use the next access time as a proxy endpoint. We have received qualitative confirmation that this variance is likely the result of overestimating LPN and CMA time in the current data pull because their

estimated time is likely to include patient wait time for the provider. We will examine this assumption once the new checkout timestamp requirement has been implemented and sufficient records have accumulated.

Appendix C – Monte Carlo Simulations

Simulations can be utilized to illustrate the effect of uncertainty in estimates when data are not available. These variations can be modeled using various probability distributions to best simulate the variations expected in the clinic. Identification of the correct models in the Crystal Ball software is necessary to provide accurate estimates in the simulations.

C.1 Crystal Ball Distributions

C.1.1 Weibull Distribution

The Weibull distribution can take on different forms depending on the shape parameter. There are three parameters in Crystal Ball for the Weibull distribution that defines the probability distribution: 1) location, 2) scale, and 3) shape. The location parameter shifts the start of the distribution where a larger location parameter shifts the start of the distribution farther away from the origin. The scale parameter affects the spread of the distribution where a larger parameter would result in a broader and shorter distribution and a smaller parameter would result in a narrower and taller distribution. While both location and scale parameter affect the “shape” of the distribution, this does not affect the functional form the distribution takes. A shape parameter is a special parameter for distributions that can take on different functional forms for the distribution. A shape parameter < 1 indicates a decreasing rate overtime, whereas a shape parameter > 1 indicates an increasing rate overtime. A shape parameter = 1 indicates no change in rate overtime.

C.1.2 Normal Distribution

A normal distribution is a symmetric bell-shaped distribution that is often used to model random variables. The normal distribution has two parameters: 1) location and 2) scale. The normal distribution is a fixed form distribution that does not have a shape parameter. The location parameter is identified by the mean and the scale parameter identified by the variance.

C.1.3 Beta PERT

A Beta PERT distribution is a smoother form of the triangular distribution defines the distribution by the minimum value, maximum value, and mode. The formula for the mean of the triangular distribution is the average of the minimum, maximum, and mode values. This distribution assumes equal weight to the minimum and maximum values indicating that these values are just as likely to occur as the value most frequent. In comparison, the formula for the mean of the Beta PERT distribution weights the mode four times the minimum and maximum values.

C.2 Simulation of Visit Cost Calculation

Monte Carlo simulations were developed to examine the effect of potential variation in minute and cost estimates. The provider time was simulated using a Weibull distribution identified by Medicare specified range for the visit type (Centers for Medicare and Medicaid Services, 2020). The scale and shape of the distribution remain constant across the delivery methods based on the distribution identified (scale = 5.5 and shape = 2.4). However, the location had to be adjusted as each delivery method had different mean provider time. The location for in-person visits was 14.65 and TH was 13.65. When controlling for aborted TH visits that were required to be seen in-person, the location was 30.55 to account for the addition of the TH labor time to the in-person visit. All other staff time was modeled on a Beta PERT distribution defined by a 10% of the estimated total time for the minimum and maximum values. Median salary for actors were varied on a normal distribution using a 10% of the mean for the standard deviation. This distribution was selected due to the narrow range often accompanying salary bands for ranks held in clinics. Simulations were then utilized to estimate total cost for the visit using the defined variation in minutes and cost per minute for actors 100,000 trials for each of the three workflows using Crystal Ball software. The distribution resulting form the simulations were fit to examine the distributions of costs.